## 3.2 Statistical Analysis Procedures

There are many different types of statistical analysis that can be performed on water quality data sets for reporting and interpretation purposes. Many inferences can be made about data from simple statistics such as mean, minimum, maximum, median, range, and standard deviation. Here is a quick review of how these statistics are calculated and how they can be used for analysis of water monitoring data. Also included in this section are some slightly more advance statistics. The following table, derived from the MPCA's Volunteer *Surface Water Monitoring Guide,* provides some guidance on the particular uses of these statistical methods.

**Table 1. Suggested Statistical Summaries for General Chemical and Physical Parameters (Adapted from *We Have Stream Data, Now What*)**

| | Statistical Summary | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Parameter** | **Average** | **Median** | **Flow-Weighted Average** | **Range** | **Quartiles** | **Confidence Intervals or Standard Deviation** | **Seasonal Average** | **Seasonal Median** | **Maximum** | **Minimum** | **Geometric Mean** |
| **Total Suspended Solids** | ■ | ■ | ■ | ■ | ■ | ■ | | | | | |
| **Temperature** | | | | ■ | ■ | | ■ | ■ | ■ | | |
| **Dissolved Oxygen** | | | | | ■ | | | ■ | | ■ | |
| **Turbidity** | | ■ | | | ■ | | | | ■ | | |
| **Nutrients** | | ■ | ■ | | ■ | ■ | ■ | | | | |
| **Conductivity** | ■ | ■ | | | ■ | | | | | | |
| **pH** | ■ | ■ | | | ■ | | | | | ■ | |
| **Alkalinity** | | ■ | | | ■ | | | | | ■ | |
| **Chlorophyll-a** | | ■ | | ■ | ■ | ■ | ■ | | ■ | ■ | |
| **Flow** | ■ | ■ | | | ■ | | | | ■ | ■ | |
| **Water Clarity/Transparency** | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | |
| **Bacteria** | | | | | ■ | | | | | | ■ |

### 3.21 Statistics

Median: The median of a data set is the middle value after all the values have been ranked in order of value. The median can easily be picked out in small data sets, or can be calculated with the =MEDIAN() equation in Microsoft Excel for large data sets.

Mean: The mean, or average, of a set of samples is one way of finding the center value of a data set. Divide the sum of the results by the number of results. Mean can be automatically calculated using the =AVERAGE() equation in Microsoft Excel.

Geometric Mean: A geometric mean can be used to calculate a mean that is not skewed by extreme values. It is one of the calculations used when assessing waters for impairment for the TMDL program, particularly for fecal coliform. Fecal coliform levels can be very low on one day and too numerous to count the next day on some streams. The geometric mean is normally close to the median for positively skewed data sets. Where G represents the geometric mean and the $x_n$ values represent a series of numbers in a data set:

$G(x_1, x_2) = \sqrt{(x_1 * x_2)} = (x_1 * x_2)^{1/2}$;
$G(x_1, x_2, x_3,) = (x_1 * x_2 * x_3)^{1/3}$;
And so on...

Note that geometric mean takes the product of all the numbers in the data set to the power of one over the number of values in the data set. Geometric mean can also be calculated automatically using a function in Excel: =GEOMEAN(A1:A5), where A1:A5 is the range of cells that contain the data to be analyzed (for the example). The geometric mean cannot be calculated for data sets that include values of zero. Therefore, values that are below the minimum detection limit (represented by <(MDL) in lab reports) must be represented by a positive number such as one-half of the MDL.

Trimmed Mean: This is another way to remove the influence of outliers in data sets. To calculate a trimmed mean, calculate the mean of only the data that falls between the 25$^{th}$ and 75$^{th}$ percentiles of a data set. Trimmed mean can be automatically calculated in Microsoft Excel by using the equation: =TRIMMEAN(). See the following section on quartiles to learn how to calculate the 25$^{th}$ and 75$^{th}$ percentiles.

Percentiles and Quartiles: Percentiles are a measure of the relative position of a single value within a data set. They are more valuable when applied to large data sets versus small ones. Percentiles are labeled $P_1$, $P_5$, $P_{25}$, etc. The subscript number refers to the percentage of the values in the data set that are smaller than the value of the percentile. So, if the $P_{30}$ percentile of a data set equals 10, 30% of the measurements are less than 10 and 70% of the measurements are greater than 10. Three particular percentiles are used quite frequently in statistical analysis. These are $P_{25}$, $P_{50}$, and $P_{75}$. These percentiles are also referred to as the 1$^{st}$, 2$^{nd}$, and 3$^{rd}$ quartiles or $Q_1$, $Q_2$, and $Q_3$, respectively. Other percentiles that are commonly used include the 5$^{th}$ and the 95$^{th}$ percentiles.

Percentiles and quartiles are another type of statistical analysis that can be performed using Microsoft Excel and other computer programs. Many programs that calculate a set of summary statistics will include the 1st, 2nd, and 3rd quartiles. To perform this calculation using a Microsoft Excel function, simply go to Insert >> Function, click on statistical, and then choose either PERCENTILE or QUARTILE. Choose the PERCENTILE function for percentiles other than the quartiles because you can input the percentile you wish to calculate (between 0 and 1). QUARTILES is a simplified version of the PERCENTILE function. The desired quartile is entered into the Quart field (0 for minimum, 1 for Q1, 2 for Q2, 3 for Q3, and 4 for maximum). Whichever function you choose, a window with two fields will appear. Enter the range of values to be analyzed into the Array field and indicate the desired percentile or quartile in the bottom field. Click OK when the information has been correctly entered into the fields.

Loads: Loads are calculated by multiplying concentration by flow volume. Daily average concentrations and/or flows can be used for continuous monitoring programs. Often, however, only one measurement for each will be available for each sampling day. Instantaneous loads can still be calculated with this data. Loads in milligrams (mg) per second (sec) can be calculated by multiplying the concentration in milligrams per liter (mg/L or ppm) by the flow in cubic feet per second ($ft^3$/sec or cfs) and then multiplying by a conversion factor of 28.31685 L/1 $ft^3$. Milligrams per day can be calculated by multiplying the mg/sec result by a conversion factor of 86,400 sec/day. After this, any other conversion factors can be applied. Kilograms per day can be calculated by multiplying the mg/day result by a conversion factor of 1 Kg/1,000,000 mg. Tons per day can be calculating by multiplying the kilograms per day by a conversion factor of 1 ton/907.1847 Kg.

Flow-Weighted Mean: Calculating the flow-weighted mean concentrations of water quality parameters places more importance to concentrations recorded during higher flows when calculating an average concentration. High flow periods can contribute the majority of the total flow volume for a given year. The concentrations of water quality parameters during periods of high flows can have a greater impact on receiving waters than the concentrations during periods of low flow. Weighted means are calculated by multiplying each individual datum in a data set by a weighting factor, finding the sum of these products, and then dividing this sum by the sum of the weighting factors. In other words, to find flow weighted mean concentrations, first multiply parameter concentration by flow for each sampling event. Find the sum of the products from all sampling events. Finally, divide this sum by the sum of all the flow values. No conversions of concentration or flow should be needed. Any conversion factors added to the equation would need to be applied to both the divisor and the dividend and will, therefore, cancel each other out and will be a waste of time. The following equation will calculate the flow weighted mean using a data set of concentrations ($c_1 \ldots c_4$) and flows ($f_1 \ldots f_4$):

$$\text{Flow weighted mean} = \frac{(c_1 * f_1 + c_2 * f_2 + c_3 * f_3 + c_4 * f_4)}{(f_1 + f_2 + f_3 + f_4)}$$

Minimum, maximum, and range: These statistics are self-explanatory. The minimum is the lowest value in the data set. The maximum is the highest value in a data set. Range is the difference between the minimum and the maximum. Minimum and maximum values can easily be found in small data sets, but equations like the MIN and MAX functions in Microsoft Excel can help find these values in a more numerous set of values in a spreadsheet.

Standard Deviation: Standard variation is a measure of the amount of variance in a data set. It is equal to the square root of the variance. This calculation can be useful in determining precision for a set of replicate samples, for example. The standard equation for standard deviation is:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(X_1 - \overline{X})^2}{n-1}}$$

In the equation above, s = standard deviation, n = the number of values in the data set; $X_1$ = the first number of the data set, $X_2$ = the second number, and so on; and $\overline{X}$ = the mean of the data set. Another way to calculate the standard deviation is shown below.

$$s = \text{the square root of } \frac{(\sum X^2 - (\sum X)^2/n)}{n-1}$$

$\sum X^2$ = Sum of the squares of the values
$\sum X$ = Sum of the values
n = Number of values

The easiest way to calculate standard deviation, however, is by using the Microsoft Excel equation: =STDEV(A1:A5), where A1:A5 is an example of a range of cells that contain the data to be analyzed.

### 3.22 QA/QC Calculations

Relative Percent Difference: Calculating the relative percent difference (RPD) between samples and duplicates can be used to measure the precision of water quality measurements. A smaller RPD indicates greater precision. Standards for RPD may be set at the beginning of a monitoring program and included in a quality assurance project plan (QAPP). Acceptable RPD standards range from <20% to <30% in existing quality assurance plans from various agencies and laboratories. The RPD between a sample and its duplicate is calculated by dividing the difference between the two samples by their average.

RPD = (Result 1 – Result 2)/[(Result 1 + Result 2)/2]*100

Percent Recovery: Percent recovery is a test of the accuracy of laboratory methods. It is essentially a ratio of the measured value versus the expected value. This test can be applied to performance evaluation sample results. Performance evaluation samples are prepared by a third party and have a known concentration. The percent recovery for a set of performance evaluation samples is equal to the measured concentration divided by the actual concentration, then multiplied by 100.

Percent recovery calculations can also be used as a method of quality control to determine if there is something in the sample or in the analytical technique that is interfering with the test. A set of duplicate samples is created from the original, real sample. A matrix spike with a known concentration of the target analyte is added to one of the duplicate samples. Both the spiked sample and the unmodified sample are analyzed at the same time. The percent recovery of a matrix spike is calculated by dividing the difference in concentration between the results for the spiked sample and the results for the original sample by the concentration of the spike that was added. Greater values for percent recovery indicate a higher level of accuracy. The lab tests a spiked sample and the non-spiked sample.

When the percent recovery is calculated, it should be within the range of 90 to 110 percent. A perfect percent recovery is 100 percent. If the percent recovery is low, there may be something in the sample that is interfering with the test. The percent recovery equation for matrix spikes is shown below.

$$\% \text{ Recovery} = \frac{(\text{Conc. of Spiked Sample} - \text{Conc. of Non-spiked Sample})}{\text{Concentration of Spike Added}} \times 100$$

### 3.23 Conversions

Conversions are often necessary when managing and analyzing water quality data. Results from different sources may be in different units. Conversions are nearly always a necessity when working with loads since the units of volume in concentration data are usually milligrams and the units of volume in flow measurements are usually cubic feet. When converting data, knowing conversion factors between units is essential. Lists of conversion factors are available in table form (see below), but they are also very handy when they are in an electronic form. Conversions can be performed with advanced calculators and with computer programs such as Convert. Convert can be downloaded for free at http://www.joshmadison.com/software/convert/.

Now that you know, for example, that one Liter equals 0.03531467 cubic feet, you still need to be able to conduct conversions based upon these conversion factors. You will need to think back to your chemistry classes. The point of a conversion is to arrive at the desired units. For example, if the average concentration of total suspended solids for a day is 50 milligrams per Liter (mg/L) and the average rate of flow for the day is 500 cubic feet per second (cfs), how many tons per day were going through the monitoring site? The desired units are tons/day. The beginning units are mg/L and ft/sec. Equations can be created in Microsoft Excel to automate these calculations, but first, write out the equation and multiply by conversion factors to cancel out units until the desired units are achieved. In this example, we want to change seconds to days, and milligrams to tons. Liters and cubic feet ($ft_3$) are both measures of volume and will be canceled out of the equation.

$$\frac{50\ mg}{1\ L} * \frac{500\ ft_3}{1\ sec} = \frac{50\ \cancel{mg}}{1\ \cancel{L}} * \frac{500\ \cancel{ft_3}}{1\ \cancel{sec}} * \frac{1\ \cancel{L}}{.03531467\ \cancel{ft_3}} * \frac{86{,}400\ \cancel{sec}}{1\ day} * \frac{1\ kg}{100{,}000\ \cancel{mg}} =$$

$$\frac{611{,}643.83\ kg}{1\ day}$$

$$\frac{611{,}643.83\ \cancel{kg}}{1\ day} * \frac{1\ ton}{907\ \cancel{kg}} = 674.36\ tons/day$$

After writing this conversion on paper, it can be translated into a Microsoft Excel equation by noting the multiplication and division factors that are applied to the original values. If the 50 mg/L is in cell A2, the 500 ft/sec value is in cell B2, and you wish to calculate the load in tons/day in cell C2, here is what the equation should look like in cell C2:

=(A2*B2*86400)/(.03531467*100000*907)

or a simplified version:

=(A2*B2*86400)/(3203040.569)

**Table 2. Useful Conversions for Water Quality Data Analysis**

| Common Conversions for the Water Quality Monitor | |
|---|---|
| **Mass** | **Area** |
| 1 gram (g) = 1000 milligrams (mg) | 1 township (twp) = 36 sections (sect) |
| 1 ton (tn) = 2000 pounds (lbs) | 1 section (sect) = 1 square mile ($mi^2$) |
| 1 kilogram (kg) = 1000 grams (g) | 1 township (twp) = 36 square miles ($mi^2$) |
| 1 kilogram (kg) = 2.20462 pounds (lbs) | 1 acre (ac) = 43,560 square feet ($ft^2$) |
| 1 pound (lb) = 453.5924 grams (g) | 1 square mile ($mi^2$) = 640 acres (ac) |
| **Distance** | 1 square mile ($mi^2$) = 2.589988 square kilometers ($km^2$) |
| 1 mile (mi) = 5280 feet (ft) | 1 square foot = 144 square inches ($in^2$) |
| 1 mile (mi) = 1.609344 kilometers (km) | 1 square meter (m2) = 10.76391 square feet ($ft^2$) |
| 1 kilometer (km) = 1000 meters (m) | 1 hectare (ha) = 2.471044 acres (ac) |
| 1 hectometer (hm) = 100 meters (m) | 1 square meter ($m^2$) = 1.19599 square yards ($yd^2$) |
| 1 meter (m) = 3.28083 feet (ft) | **Computer Terminology** |
| 1 meter (m) = 39.36996 inches (in) | 1 kilobyte (KB) = 1024 bytes |
| 1 meter (m) = 100 centimeters (cm) | 1 megabyte (MB) = 1024 kilobytes (KB) |
| 1 centimeter (cm) = 10 millimeters (mm) | 1 gigabyte (GB) = 1024 megabytes (MB) |
| 1 meter (m) = 1.09361 yards (yd) | **Pressure** |
| 1 yard (yd) = 3 feet (ft) | 1 inch of mercury = 25.4 millimeters of mercury |
| 1 inch (in) = 25.4 millimeters (mm) | 1 inch of mercury = 3.386388 kilopascals (kPa) |
| **Time** | 1 inch of mercury = 33.86388 millibars (mb) |
| 1 year (yr) = 365 days | **Volume** |
| 1 day = 24 hours (hrs) | 1 liter (L) = 1000 milliliters (ml) |
| 1 hour (hr) = 60 minutes (min) | 1 cubic foot ($ft^3$) = 28.31685 liters (L) |
| 1 minute (min) = 60 seconds (sec) | 1 gallon = 3.785412 liters (L) |
| 1 hour (hr) = 3600 seconds (sec) | 1 liter (L) = 33.81402 ounces (oz) |
| 1 day = 86,400 seconds (sec) | 1 cubic yard (yd3) = 27 cubic feet |
| **Flow** | **Concentration** |
| 1 cubic foot/second (cfs) = 646316.9 gallons/day = 2446576 liters/day = 101940.6 liters/day = 2446.576 cubic meters/day = 3600 cubic feet/hour | 1 milligram/liter (mg/L) = 1 part per million (ppm) = 1000 micrograms/liter (μg/L) |
| | 1 microgram/Liter (μg/L) = 1 part per billion (ppb) |
| **Temperature** | |
| Fahrenheit to Celsius: C = (F-32) * 5/9 (Subtract 32, multiply by 5, and then divide by 9.) | |
| Celsius to Fahrenheit: F = 32 + C * 9/5 (Multiply by 9, divide by 5, and then add 32) | |
| **Miscellaneous Conversions** | |
| 1 cubic yard of sediment = about 2,500 pounds or 1.25 tons | |
| Amount of sediment in a two-axle, 5 yard dump truck load = 6.25 tons | |
| Amount of sediment in a tri-axle, 12 yard dump truck load = 15 tons | |

### 3.24 Graphical Methods

Other forms of statistical analysis are often needed. Summarizing analysis results in tables, graphs, or charts for reporting purposes can be very helpful to the reader. Some of the descriptive statistical analysis performed for the Red River Watershed Assessment Protocol Project include the determination of minimum detection limits, recommending methods for addressing values below the minimum detection limit, histograms, boxplots, time series plots (next section), correlation matrixes, and flow duration curves. It is important to make graphs neat, informative, and understandable. The graphs should be useful for interpreting the meaning of data and presenting findings from data. There are many techniques involved in creating quality graphs. Here are some tips:

- ✓ Graphing data is part of a process. You may end up graphing more data than you will use in a report or presentation. Some data you graph will be more valuable than others. If graphs are used as part of the process of understanding data, their meanings, indications, and other results may be summarized in another form and the graphs may not necessarily appear in the final report or presentation.

- ✓ Column graphs should be used with discrete data (data that is not continuous). Line graphs are used with continuous data. Line graphs that are used for discrete measurements may mislead the viewer into thinking the data is continuous. An example of a good line graph would be flow data that is collected at regular intervals (hourly, every 15 minutes).

- ✓ Have a clear title.

- ✓ Make sure you have simple clear label on the axes that shows reporting limits.

- ✓ Use a scale size that reveals trends, adjust it from the default scale to meet your needs.

- ✓ Avoid clutter.

- ✓ Illustrate information that allows the reader to get to the point quickly. Use graphs only when they convey meaningful information.

- ✓ When displaying data from multiple sites, displaying information from upstream to downstream is an intuitive way to organize and present your results.

- ✓ Consider the background and graph colors. Do they print well? Adjust colors to create a color scheme that will make sense to the reader.

- ✓ Just do it! Start in and play around with different types of graphs…thankfully, there is an undo button.

Histogram/Frequency Plot: Histograms and frequency plots show the distribution of observations within a sample set. They are usually used to visually assess the degree of scatter and whether the observations are normally distributed. Meaning, if the observations are normally distributed, the heights of the columns should be roughly shaped like the Normal distribution curve (the superimposed blue line in the example below. These graphs can be used to interpret the symmetry and variability of data. Symmetric data will be structured symmetrically around a central point. The extent and direction to which data is being skewed will also be indicated by boxplots and frequency distributions.

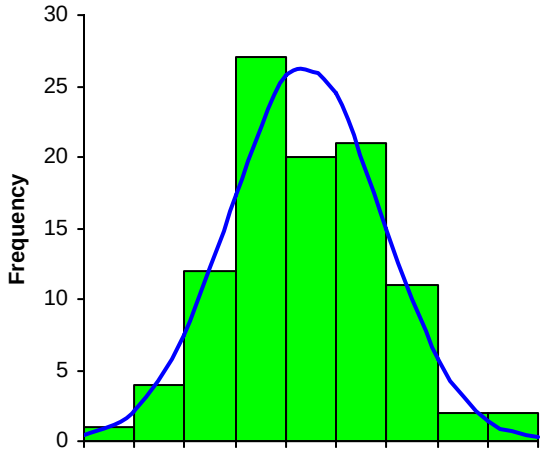



**Figure 12. Example Frequency Plot**

Both histograms and frequency plots split data into intervals, count the number of values in each group, and displaying the data in the form of a bar chart (green bars in Figure 12). There are two differences between the two graphs. The vertical axis of a histogram represents the percentage of the total data set that is included in each interval. The vertical axis of a frequency plot represents the number of observations within an interval. These plots can either be created manually (see example in figure 2) or using a computer program. Analyse-it, an add-in for Microsoft Excel ($100), histogram creating add-ins for Microsoft Excel (around $30), the (free) data analysis add-in for Microsoft Excel, and StatCrunch (free online at http://www.statcrunch.com/) are some of the programs that can be used to create histograms.

The Webstat/StatCrunch program is an online statistical analysis tool that can be accessed through the RLWD website on the Analyze or Download Data page for each monitoring site. To get to this page, go to the RLWD website at www.redlakewatershed.org, click on the Water Quality section, search for a site using the interactive map or text search tools, click on a blue site ID number (the link to the informational pages for the monitoring site), and then click on the Analyze or Download Data tab. Scroll down to the blue link for the current version of StatCrunch.

After you have created an (free) account, the software will automatically load the data from the monitoring site into the program. The data can then be analyzed using nearly

any type of applicable statistical or graphical analysis. The statistics available in StatCrunch include correlation, covariance, summary statistics for columns or rows, frequency tables, contingency tables, z statistics, proportions, variance, regression, t statistics, ANOVA, and control charts. The options available in StatCrunch for graphical analysis include bar plots, pie charts, histograms, stem and leaf plots, boxplots, dot plots, means plots, QQ plots, scatter plots, index plots, chart group statistics, parallel coordinates, pairs plots, 3D rotating plots, and color schemes.

Consider the following 22 samples of a contaminant concentration (in ppm): 17.7, 17.4, 22.8, 35.5, 28.6, 17.2 19.1, <4, 7.2, <4, 15.2, 14.7, 14.9, 10.9, 12.4, 12.4, 11.6, 14.7, 10.2, 5.2, 16.5, and 8.9.

STEP 1: This data spans 0 - 40 ppm. Equally sized intervals of 5 ppm will be used: 0 - 5 ppm; 5 - 10 ppm; etc. The endpoint convention will be that values are placed in the highest interval containing the value. For example, a value of 5 ppm will be placed in the interval 5 - 10 ppm instead of 0 - 5 ppm.

STEP 2: The table below shows the number of observations within each interval defined in Step 1.

STEP 3: The horizontal axis for the data is from 0 to 40 ppm. The vertical axis for the frequency plot is from 0 - 10 and the vertical axis for the histogram is from 0% - 10%.

STEP 4: There are 22 observations total, so the number observations shown in the table below will be divided by 22. The results are shown in column 3 of the table below.

STEP 5: A common unit for this data is 1 ppm. In each interval there are 5 common units so the percentage of observations (column 3 of the table below) should be divided by 5 (column 4).

STEP 6: The frequency plot is shown in Figure 2-1 and the histogram is shown in Figure 2-2.

| Interval | # of Obs in Interval | % of Obs in Interval | % of Obs per ppm |
|---|---|---|---|
| 0 - 5 ppm | 2 | 9.10 | 1.8 |
| 5 - 10 ppm | 3 | 13.60 | 2.7 |
| 10 - 15 ppm | 8 | 36.36 | 7.3 |
| 15 - 20 ppm | 6 | 27.27 | 5.5 |
| 20 - 25 ppm | 1 | 4.55 | 0.9 |
| 25 - 30 ppm | 1 | 4.55 | 0.9 |
| 30 - 35 ppm | 0 | 0.00 | 0.0 |
| 35 - 40 ppm | 1 | 4.55 | 0.9 |

**Figure 13. Example of Generating a Histogram and a Frequency Plot.**

The most common available option for the creation of a histogram within a spreadsheet is likely to be the data analysis add-in for Microsoft Excel. Before starting, you will need to create a column of values that will specify the borders of the intervals within the histogram you will be creating. To see if this add-in is loaded in your version of Excel, click on **Tools** menu. If you do not see **Data Analysis** in the **Tools** menu, click on **Add-ins** instead. A window will appear that shows a list of possible add-ins for Excel. Check the box for **Analysis ToolPak** and click **OK** to install the add-in. You will likely need to insert your Microsoft Office CD in order to complete the installation. Once the installation process is complete, you can open the data analysis window by clicking on **Data Analysis** in the **Tools** menu. Within this window, you can see all the different types of statistical analysis that can be performed with this tool. To create a histogram, double click on **Histogram** in the list of options. The histogram window will then appear. In this window, you will need to specify the **input range**. This is the set of values you want to

analyze. The **BIN range** is the column of numbers that you created at the beginning of these instructions. Indicate where you want the histogram to appear by specifying an output range or by telling the program to create a new worksheet. Check the **chart output** box to get a bar chart histogram. When you click OK, the program will create the histogram.

Boxplots: Creating boxplots (or box and whisker plots) is another method for visually representing the distributions within a data set. Boxplots show the relative positions of Q1, Q2, Q3, minimum, and maximum are shown above a scaled real number line. The minimum and maximum values of the data set are represented by lines drawn from the ends of the box. The left side of the box represents Q1, the first quartile. 25% of the samples are less than the value of Q1. Q3 is represented by the right side of the box and Q2 is represented by a line drawn in the middle of the box. They can be used to compare sites by placing a boxplot for each site on the same graph. Box and whisker plots can also be used to determine if sites are even comparable. If the boxes of two sites do not overlap, the sites are not comparable. This is because the best water quality of one site at its best is almost always worse than the water quality of the other site at its worst.

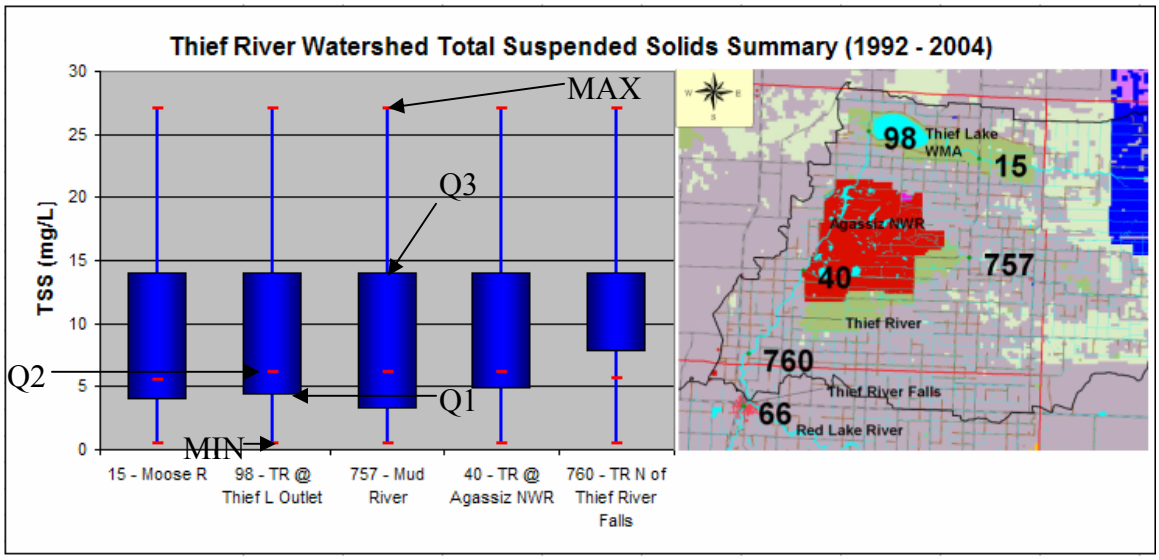



**Figure 14. Boxplot of TSS results within the Thief River Watershed with map.**

Several different methods for generating boxplots and histograms using software have been used by the RLWD. One of these is the Analyse-It software that can be purchased for approximately $100 as an add-on for Microsoft Excel. Existing Excel data can easily be used for the calculation of "over 30 parametric & non-parametric statistics, including descriptive statistics, box-whisker plots, correlation, multiple linear regression analysis, ANOVA, & chi-square statistics." This program basically creates a worksheet that is set up as a report and includes histograms, percentiles, and summary statistics along with the boxplots. Another way to create boxplots, along with nearly any type of statistical analysis can be performed, is by using the Webstat/StatCrunch program.

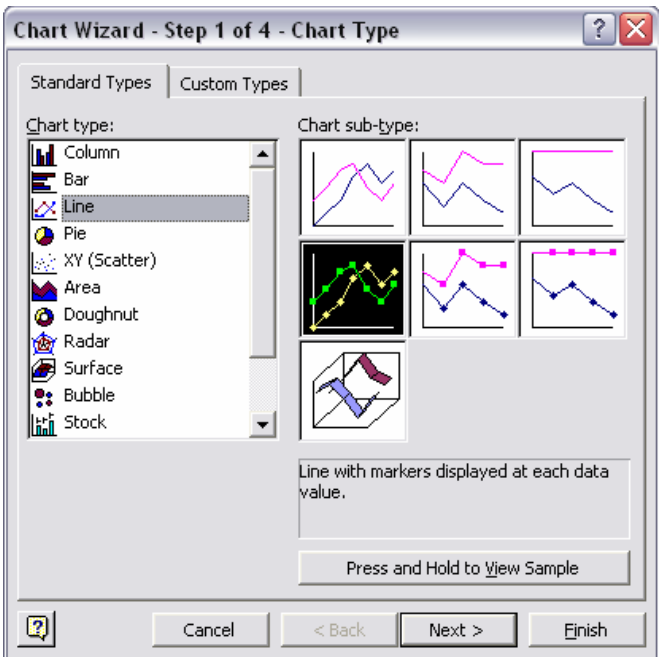
The preceding methods definitely work, but a user sometimes may want a worksheet dedicated to boxplots. In this case, boxplots can be created using the Chart Wizard in Microsoft Excel. Since there is no preset setting (as of Office 2000) for boxplots, the program needs to be tricked into creating a boxplot. The following step-by-step methods expound upon those found in *We Have Data, Now What?,* a manual compiled for the Data Analysis and Interpretation Pilot Training Workshop for Citizen Volunteer Water Quality Monitoring Programs workshop by the Red River Basin Monitoring Network, Rivers Council of Minnesota, and the River Network.

1. The first step to creating a box and whisker plot, or boxplot, is to determine which monitoring sites will be featured on the graph and create the summary statistics that will be used to create the plot. In the summary statistics table, sites should be placed in a significant order, such as upstream to downstream. The summary statistics necessary for creating a boxplot are the **25th percentile (Q1), minimum, median (50th percentile or Q2), maximum, and the 75th percentile (Q3)**. If the columns are in this order, as shown below, you will be able to skip Step 13. Also, after saving the boxplot as a custom chart type, having summary data arranged in this order will make the creation of boxplots easier in the future.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | Fecal Coliform | | | | | | |
| 2 | | 25th% | min | median | max | 75th% | avg | # of samples |
| 3 | CR23 | 30.0 | 4.0 | 46.0 | 186.0 | 57.8 | 56.2 | 20 |
| 4 | SG130 | 20.0 | 2.0 | 60.0 | 314.0 | 74.0 | 59.2 | 21 |

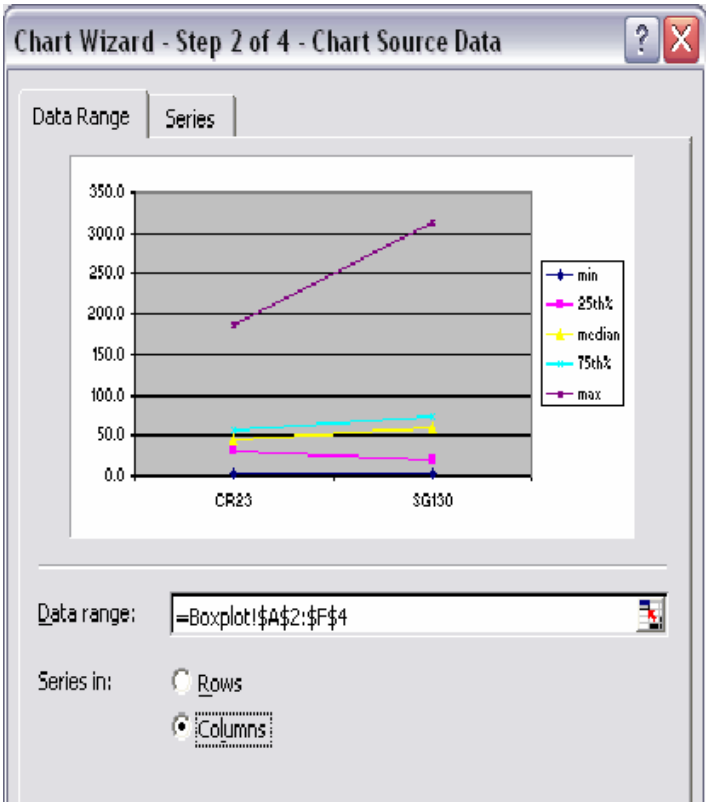
2. **Select** the site name, 25th percentile, minimum, median, maximum, and 75th percentile **column headings and data**.
3. Select the **Chart Wizard Button**. 

4. In the **Chart Wizard Step 1 of 4**, click on the **Standard Types** tab and choose the **Line** chart. Choose the chart sub-type labeled **"line with markers displayed at each data value."**



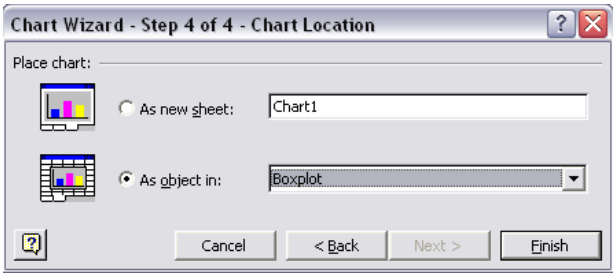

5. Click **Next** to continue.

6. In **Chart Wizard Step 2 of 4**, the data range box should automatically contain the summary data cells you selected in Step 2. Click the round button that puts the series into **Columns**. Click **Next** to continue.

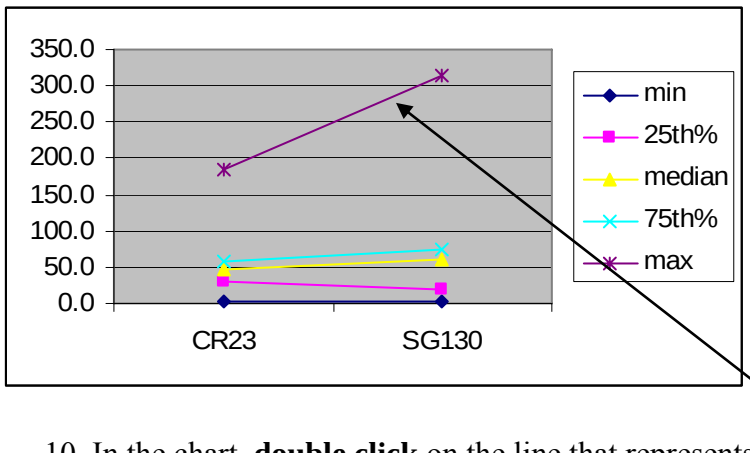


7. **Skip** the **Chart Wizard Step 3 of 4** for now by clicking **Next** to continue.

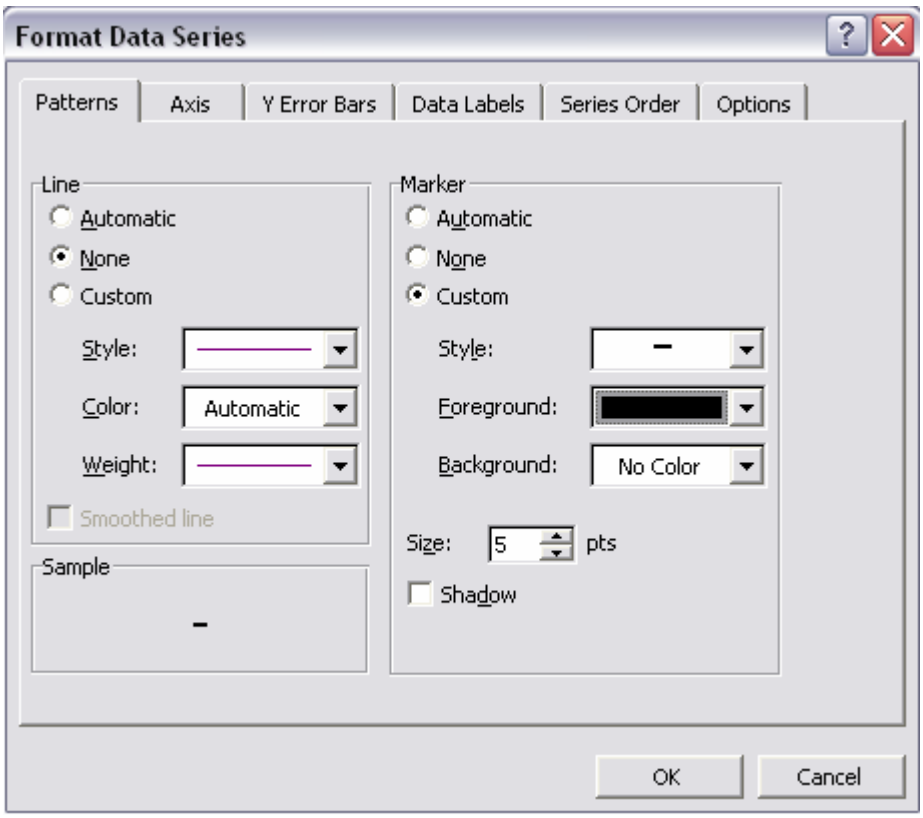
8. In the **Chart Wizard Step 4 of 4 – Chart Location**, you can choose the location of the graph. It can either be placed in its own worksheet, or in another worksheet that, for example, is dedicated to graphic analysis.



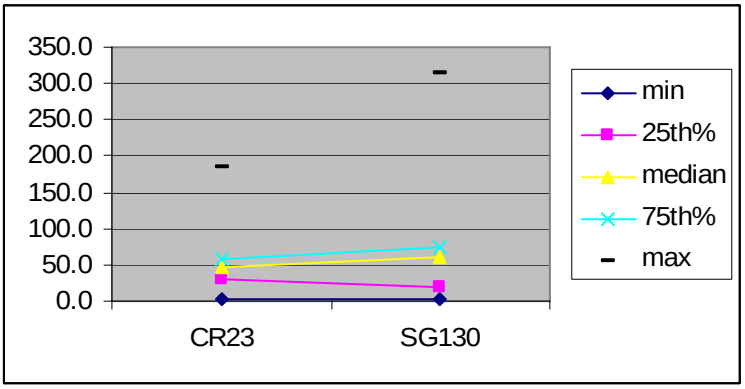

9. Now you have the beginnings of a chart that should look something like the one below. You may need to adjust the scale and fonts to make sure the chart is readable. This and other aspects of the appearance can also be adjusted when the chart is completed so it is not necessary at this point.



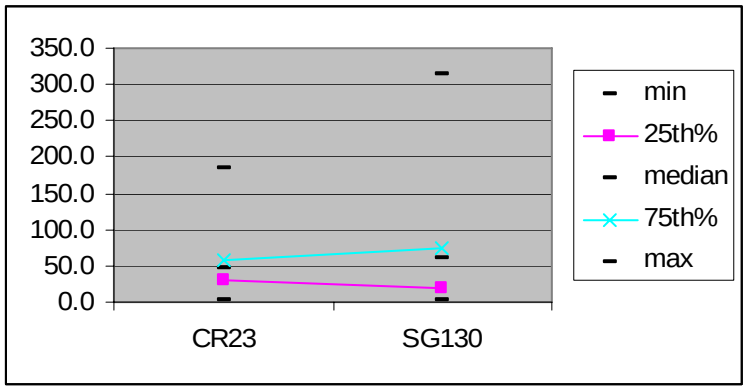

10. In the chart, **double click** on the line that represents the **maximum** values in the data set. In the **Patterns** tab, remove the line by choosing **None** under **Line**, change the **Marker Style** to a dash (-), and change the **Marker Foreground Color** to **black**.



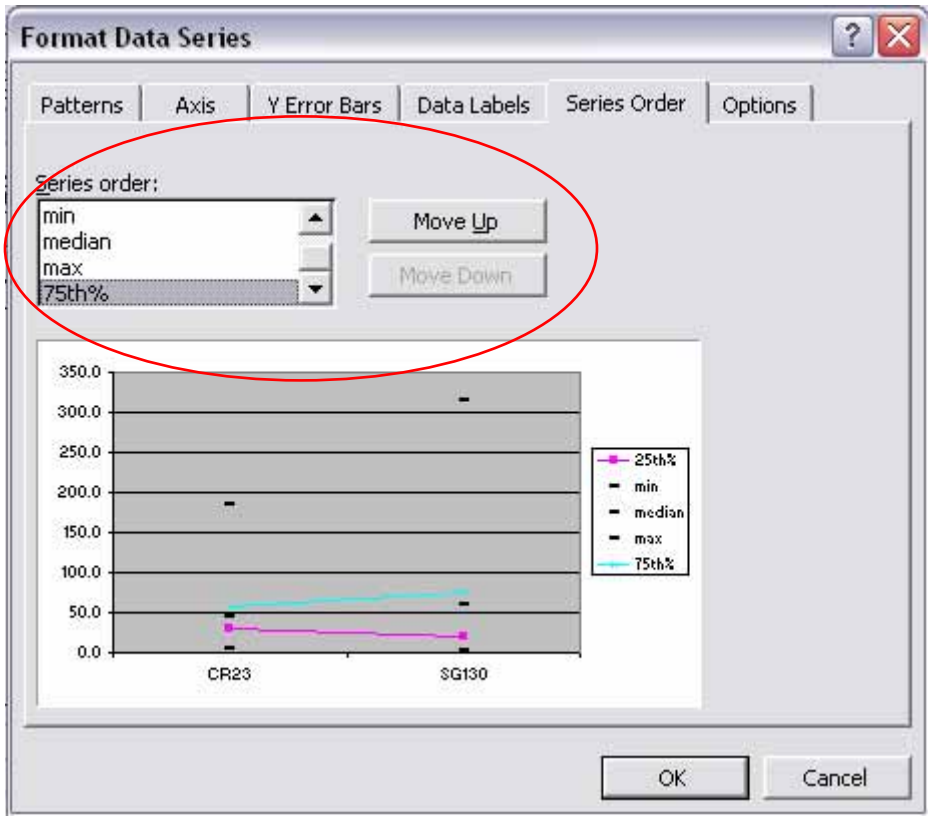

11. Now the graph should look similar to this:



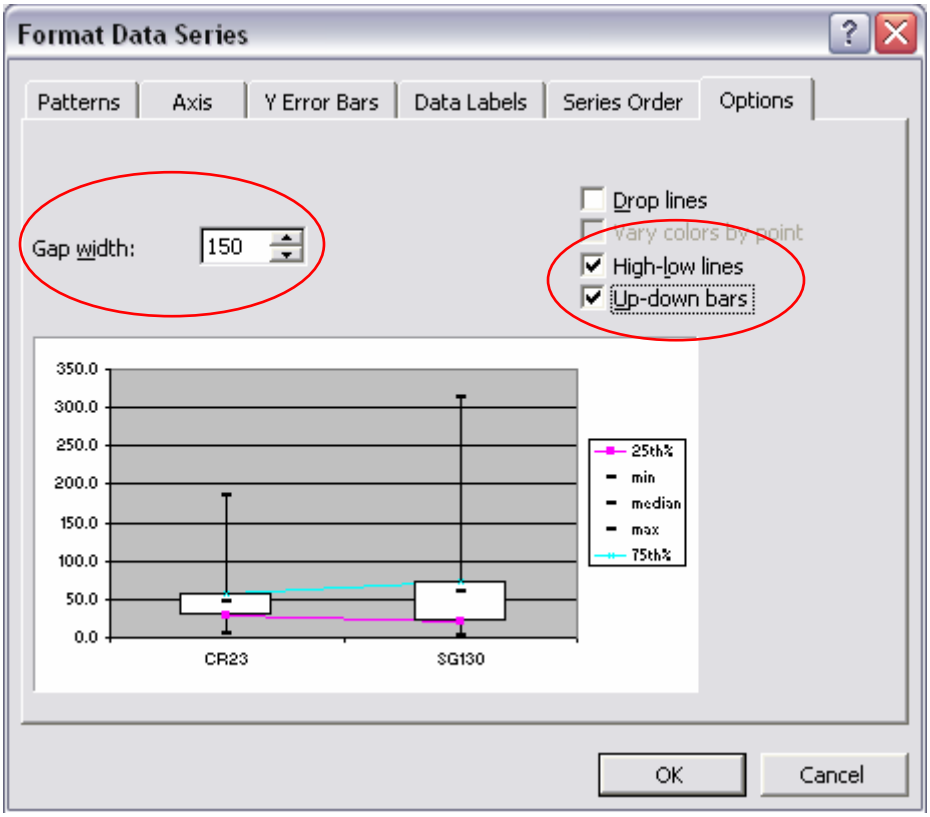
12. **Repeat Step 10** for the **minimum** and **median** lines. When you are done, the graph should look like this:



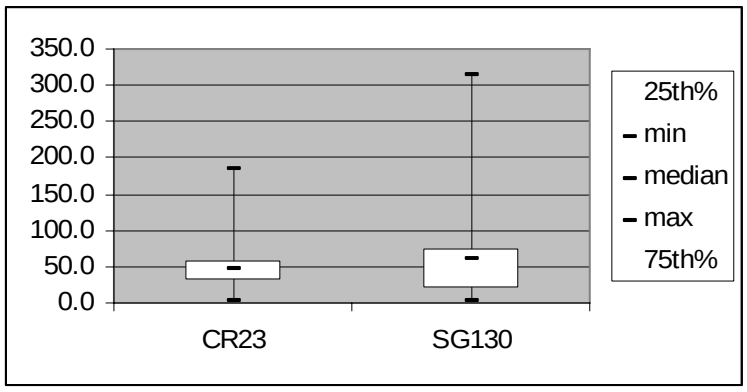

13. **Double-click** on the line for the **25th** or **75th** data series to bring up the **Format Data Series** window. This time, select the **Series Order** tab. Make sure that the order of the series to the following: **25th percentile, minimum, median, maximum, 75th percentile**. This series order can be changed, if needed, by using the **MOVE UP** and **MOVE DOWN** keys.



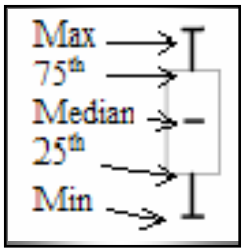
14. **Before clicking OK, click** the **Options** tab. **Check** the boxes for **High-low lines** and **Up-down bars**. Adjust the **Gap width** number to **150**. A smaller gap width value will produce larger boxes in the box and whisker plot, and vice-versa.




15. Click on the **Patterns** tab and **repeat Step 10** for the **25th** and **75th** percentile lines to **remove** the remaining **lines** and **markers**. Now the graph should look similar to this:

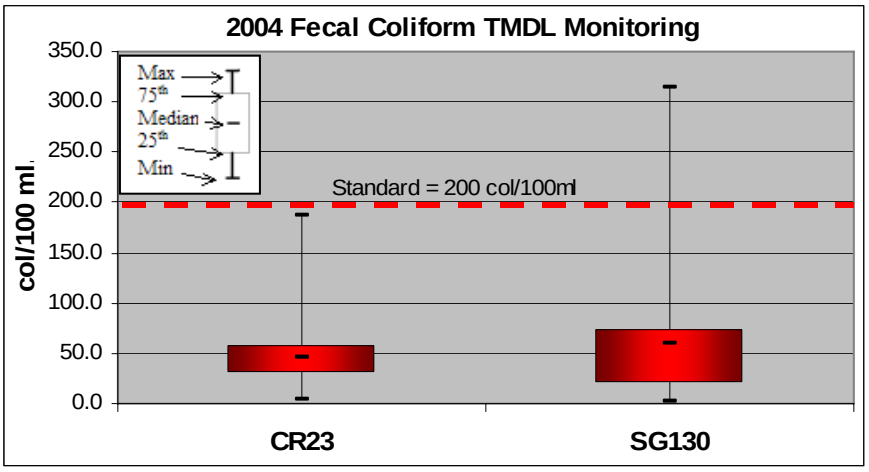


16. Now you can begin to format the appearance of the chart. You can double-click on the boxes to bring up the **Format Up** Bars window and change their color, add shading, etc. Remove the legend and make your own (like the one below). Excel doesn't seem to have a legend that works for these graphs.



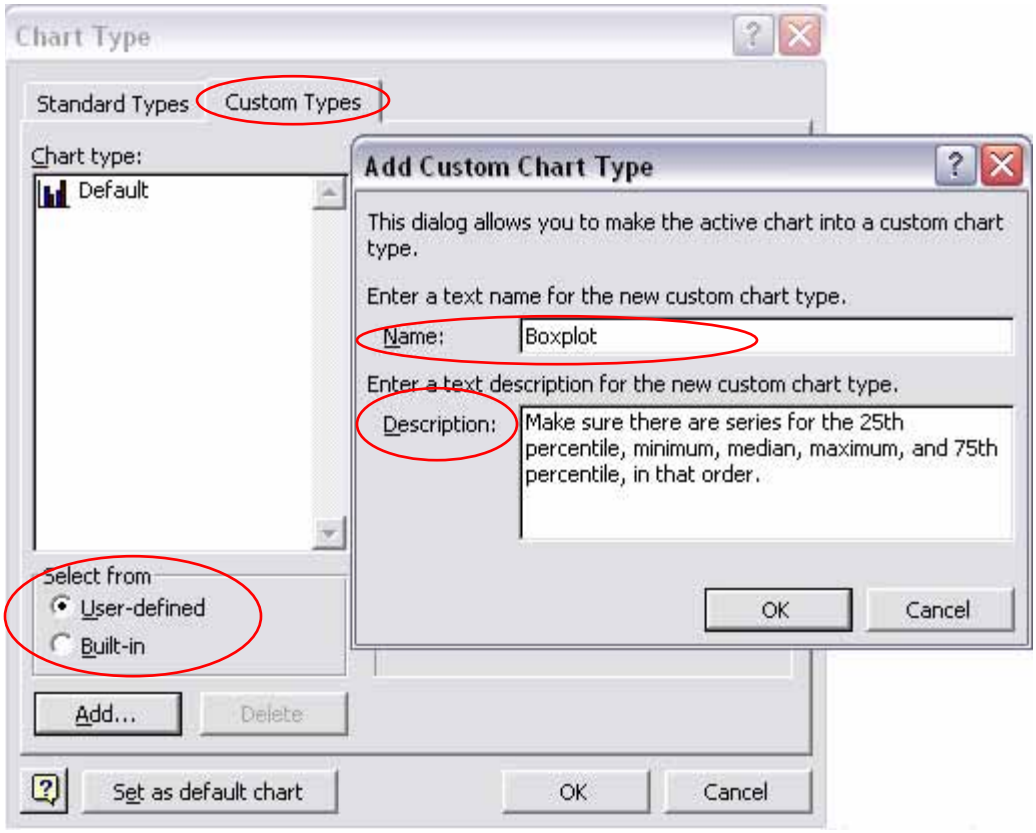

17. To change the scale or fonts, double-click on those specific parts of the graph (such as the site names on the X axis or the numeric values on the Y axis) to open the Format Axis window and change the formatting, scale, or font size.

18. To add a title, go to the Chart➔Chart Options➔Title and fill in the appropriate title. Also, lines can be added to the chart to indicate water quality standards. The final box and whisker may look like this:




Note: If there is a large degree of difference between the sites you may want to adjust the scale to show the sites that are "crunched up" in a small data range. You could also remove the sites.

19. After completing the box and whisker plot, save the style so that you can skip steps 1-12 the next time you want to create a box and whisker graph. Do save the style, right click on the chart and select **Chart Type**. Click on the **Custom Types** tab. Select the **User-defined** button. Click the **Add button.** The **Add Custom Chart Type** window will appear. Name the new custom type "**Boxplot"** or "**Box and Whisker"** and type a description. The necessary series order is an important piece of information to put in the description box. Click **OK** when you are done. An option for creating boxplots will appear among the chart type options.




If a custom chart type has been created for box and whisker plots, additional boxplots can be made very easily and efficiently. Some of the steps in the process can be skipped. To create a box and whisker plot using the custom chart type that was created in step 19, first complete Steps 1 – 3. Instead of choosing the chart type indicated in Step 4, choose the custom chart type created in Step 19: Chart Wizard➔Custom Types➔User-defined➔(Name of custom box and whisker plot chart type). Perform steps 5-8, and then skip to step 13. If your columns were in the correct order ($25^{th}$➔min➔median➔max➔$75^{th}$), step 13 is also unnecessary and can be skipped. For step 14, look at the preview of the chart under the Format Data Series➔Options tab to determine whether or not you need to adjust the gap width. Step 15 and 19 can be skipped, but steps 16-18 are still needed in order to adjust the appearance of the graph, add a title, etc. The following page shows what Steps 4-6 will look like when using the custom chart type for boxplots (created in Step 19).

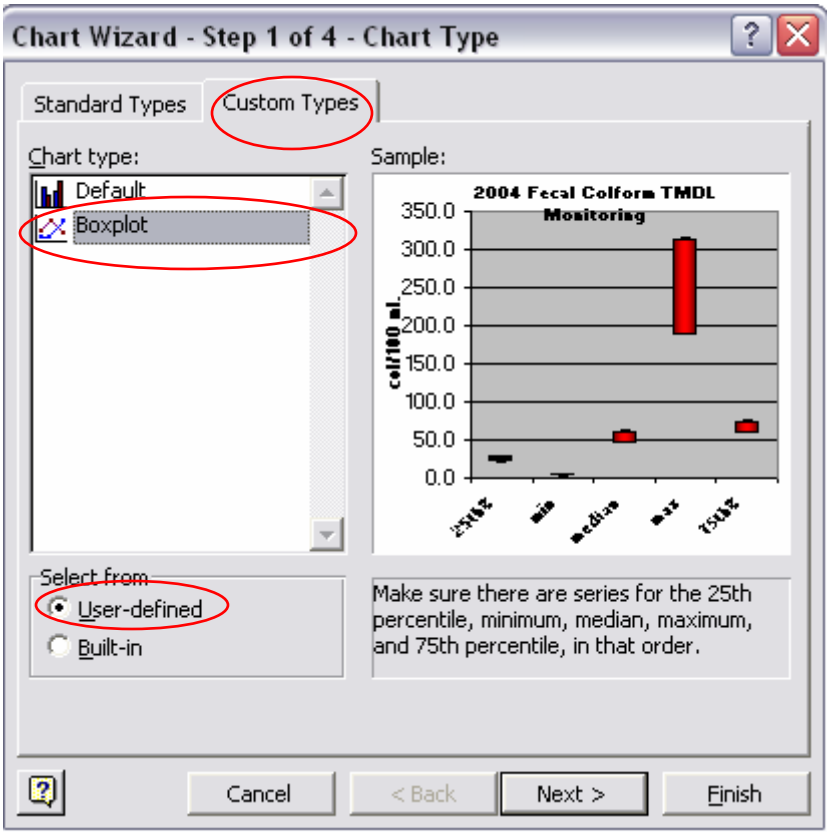


Chart Wizard - Step 1 of 4 - Chart Type
Standard Types
Custom Types
Chart type:
Default
Boxplot
Sample:
2004 Fecal Colform TMDL Monitoring
350.0
300.0
250.0
200.0
150.0
100.0
50.0
0.0
Select from
User-defined
Built-in
Make sure there are series for the 25th percentile, minimum, median, maximum, and 75th percentile, in that order.
Cancel
< Back
Next >
Finish

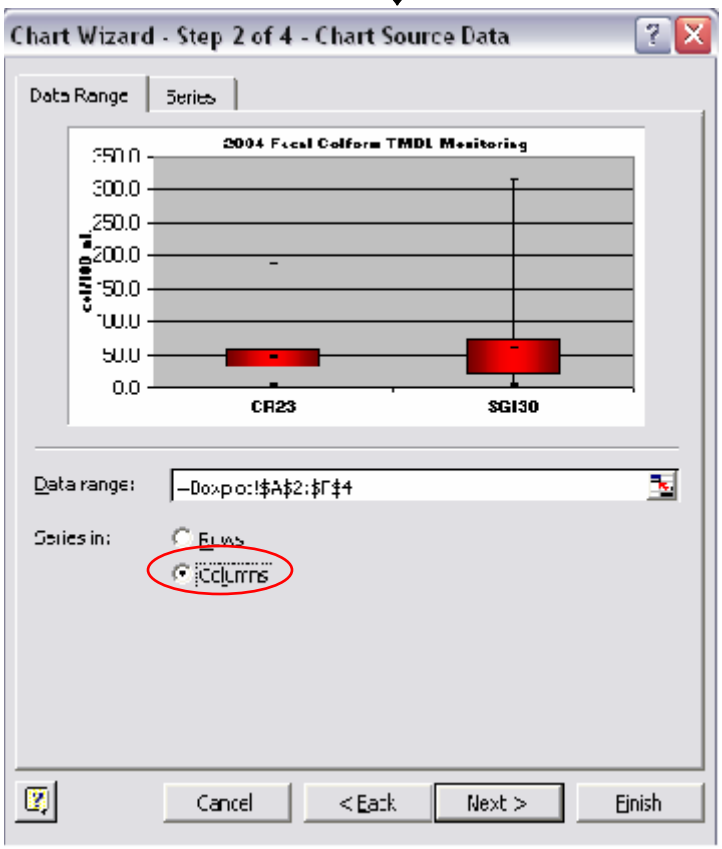


Chart Wizard - Step 2 of 4 - Chart Source Data
Data Range
Series
2004 Fecal Colform TMDL Monitoring
300.0
250.0
200.0
50.0
0.0
CR23
SGI30
Data range:
Series in:
Columns
Cancel
< Back
Next >
Finish

## 3.25 Measures of Association

Correlation matrixes, Pearson's correlation coefficient, Spearman's rank correlation coefficient and serial correlation coefficient are all measures of association in data sets. In other words, the purpose of determining correlation is to tell how closely x and y values are related (i.e. water temperature and dissolved oxygen or turbidity and total suspended solids).

Correlation matrixes are a graphical method of determining correlation. In Microsoft Excel, x values can be plotted against y values in a scatter plot. This scatter plot can be created using methods similar to those described in section 2.3. A time series plot may be considered a correlation matrix of comparing water quality data to time. This can be used as a quick way to determine correlation between two sets of data. The difference between time series plots and correlation plots is that the data points are not chronological on correlation matrixes and correlation matrixes can have parameters on both the x and the y axis instead of just on the y axis.

In Microsoft Excel, a trendline can be added to the data plot by right clicking on the data points and selecting "**Add Trendline**" and checking the "Display $R^2$" box under the "Options" tab in the **Add Trendline** window. A user can visually assess how well the plotted points are clustered along the trendline and by observing the $R^2$ value. The $R^2$ value also shows how reliably the equation of the trendline can be used to predict y values based on x values. It is the square of the correlation coefficient. An $R^2$ value that is close to 1 indicates a close association between x and y values.

Since not all trends are linear, using a trendline in Excel gives the user the advantage of being able to create polynomial, exponential, logarithmic, and moving average trendlines. When reporting results from trend analysis, creating a summary table of trend analysis results may be preferable to pages and pages of correlation matrix graphs.

Plotting correlation matrixes is very helpful, but not always necessary. Direct calculation of a correlation coefficient may be a desirable alternative for measuring the amount of association between two sets of data. Correlation matrixes can be used to find relationships between turbidity and total suspended solids, turbidity and transparency tube readings, water temperature and dissolved oxygen, turbidity and dissolved oxygen, turbidity (or total suspended solids) and phosphorus, flow and temperature, flow and dissolved oxygen, or other parameter combinations.

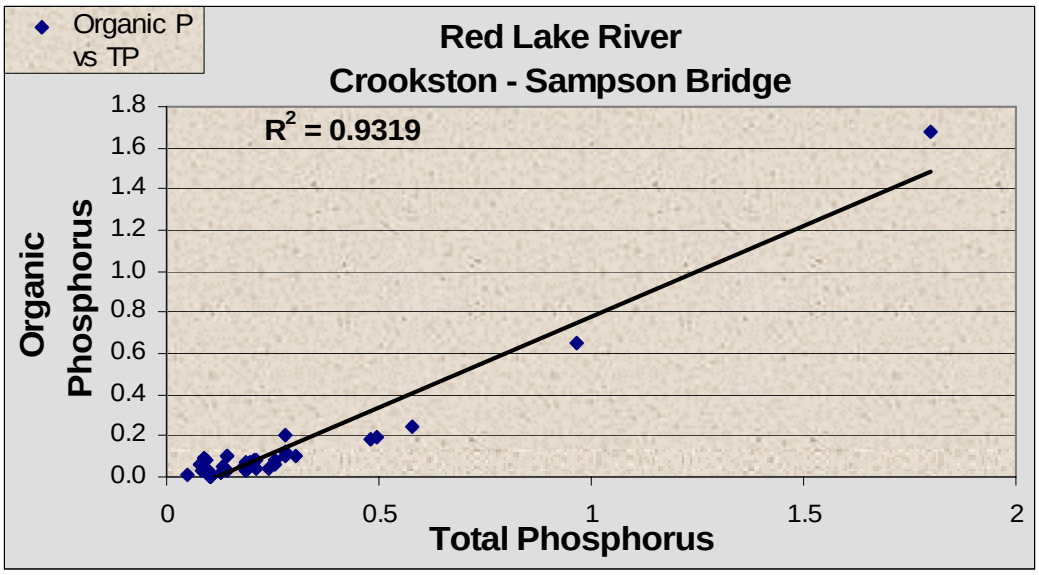



**Figure 15. Example of a Correlation Matrix**

Regression: Regression, as a statistic, can be used to find a relationship between two variables and to estimate the value of one variable based upon the value of another. Finding a relationship between two variables using regression is particularly useful because, especially in water quality monitoring, rarely, if ever, is there a direct mathematical relationship between variables. Although linear regression can be calculated and plotted by hand using the equations and methods found in textbooks, the goal of this document is to increase efficiency in data analysis. Therefore, the use of Microsoft Excel for the creation of scatter plots and trendlines is recommended. In Excel, a trendline (regression line) can easily added to a scatter plot. Sections 2.25 and 2.31 give further instructions for creating and analyzing xy scatter plots in Excel. The equation (including the slope) and the $R^2$ (coefficient of determination) value for the line can be displayed on the graph as well.

Pearson's product-moment correlation coefficient: This is a commonly used method of correlation analysis that measures a linear relationship between two variables. Possible values for the Pearson's correlation coefficient range from -1 to 1. Negative values signify a negative slope and positive values signify a positive slope. A value of -1 represents a perfectly negative linear correlation. A value of +1 indicates a perfectly positive linear correlation. Values close to 0 indicate very little correlation between the two variables. The closer the correlation coefficient is to -1 or +1, or the closer its square is to 1, the more correlation there is between the two variables. The Pearson's correlation coefficient is calculated using the equation shown in the figure below, taken from the EPA's *Guidance for Data Quality Assessment –Practical Methods for Data Analysis, EPA QA/G-9*.

It can also be calculated using the Microsoft Excel equation: =PEARSON(,). To insert this function into a cell, go to Insert>Function, highlight the statistical category of available functions, and then double-click PEARSON or highlight it and click OK. A box will then appear that will ask for the two data sets that will be analyzed for correlation (array 1 and array 2). Excel also has a CORREL(,) function for calculating a correlation coefficient.

Box 2-6: Directions for Calculating Pearson's Correlation Coefficient with an Example

Let $X_1, X_2, \ldots, X_n$ represent one variable of the n data points and let $Y_1, Y_2, \ldots, Y_n$ represent a second variable of the n data points. The Pearson correlation coefficient, r, between X and Y is computed by:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\left[\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}\right]\right]^{1/2}}$$

Example: Consider the following data set (in ppb): Sample 1 — arsenic (X) = 8.0, lead (Y) = 8.0; Sample 2 - arsenic = 6.0, lead = 7.0; Sample 3 - arsenic = 2.0, lead = 7.0; and Sample 4 - arsenic = 1.0, lead = 6.0.

$$\sum_{i=1}^{n} X_i = 10, \quad \sum_{i=1}^{n} Y_i = 28, \quad \sum_{i=1}^{n} X_i^2 = 105, \quad \sum_{i=1}^{n} Y_i^2 = 198, \quad \sum_{i=1}^{n} X_i Y_i = (8x8) + \square\ldots + \square(1x6) = 126.$$

and $$r = \frac{126 - \frac{(17)(28)}{4}}{\left[\left[105 - \frac{(17)(17)}{4}\right]\left[198 - \frac{(28)(28)}{4}\right]\right]^{1/2}} = 0.865$$

Since r is close to 1, there is a strong linear relationship between these two contaminants.

**Figure 16. Equations and Directions for Calculating Pearson's Correlation Coefficient by Hand**

Spearman's correlation is a method for calculating correlation coefficient that is less sensitive to extreme values than the Pearson's correlation coefficient and is not affected by transformed data. For this method, the same equation is used for calculating the coefficient as the Pearson's coefficient, but there is a data transformation involved. The values for each variable are changed to their rank within their respective data sets. This is relatively simple to do in Microsoft Excel. New columns can be added to a spreadsheet next to each column of raw or transformed data that is going to be used for the correlation analysis. Input the rank of each value into its respective new column (Hint: the Data>Sort function and the sort ascending ( ) button are useful for this task). Once the ranks have been entered, the correlation efficient is determined for each variable's ranking data. If there is not a good statistical relationship between each variable (Pearson's coefficient), this type of correlation analysis will determine if larger values of x correlate with larger values of y and smaller values of x correlate with smaller values of y.

For example, the Pearson's correlation coefficient calculated to determine the correlation between total suspended solids and flow at site #760 on the Thief River was only .27. This indicates that there is not a strong relationship between the two variables. However, the Spearman's method resulted in a correlation coefficient of .74, which indicates a stronger relationship than the Pearson's correlation coefficient. This tells us that higher flows at the monitoring site may be related to higher levels of total suspended solids, even though there is not a linear relationship between the two parameters.

Using a correlation matrix to identify and remove outliers can help increase any correlation coefficient. This affects the Pearson's correlation coefficient more than it affects the Spearman's correlation coefficient, since the Spearman's coefficient is affected less by extreme values. After removing only two outliers in the site #760 TSS vs. flow data set, the Pearson's correlation coefficient increased from .27 to .55, while the Spearman's correlation coefficient only increased to .74 from .76. Since a data set with nearly zero correlation can be made to look like one with a good correlation if enough outlying data is removed, the practice of removing a large number of outliers in order to improve correlation plots is not encouraged. Instead, analysis for association using the Spearman's correlation coefficient, transformation of data to natural log values, or using polynomial trendlines in Microsoft Excel may be used if a correlation is not found with other methods.

### 3.26 Pivot Tables

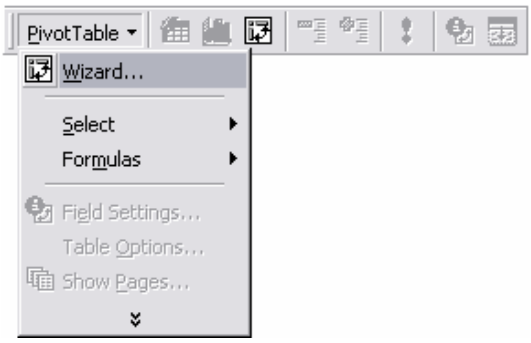
The user guide for Microsoft Excel describes a pivot table as "an interactive worksheet table that quickly summarizes large amounts of data using a format and calculation methods you choose. It is called a pivot table because you can rotate its row and column headings around the core data area to give you different views of the source data." (sic). They are useful for summarizing large amounts of data, such as continuous monitoring data, from which daily averages can be calculated from hourly data by creating a pivot table. Tables can be created that summarize a data set using sum, average, maximum, minimum, standard deviation, variance, count, or product calculations. The following is a set of step-by-step directions that show how to create a basic pivot table. Although menu composition, precise methods, and window appearance may vary among different versions of Microsoft Excel, the basic process for creating the tables should be the same.

1) Open an Excel file that contains a worksheet with the raw data you wish to analyze.

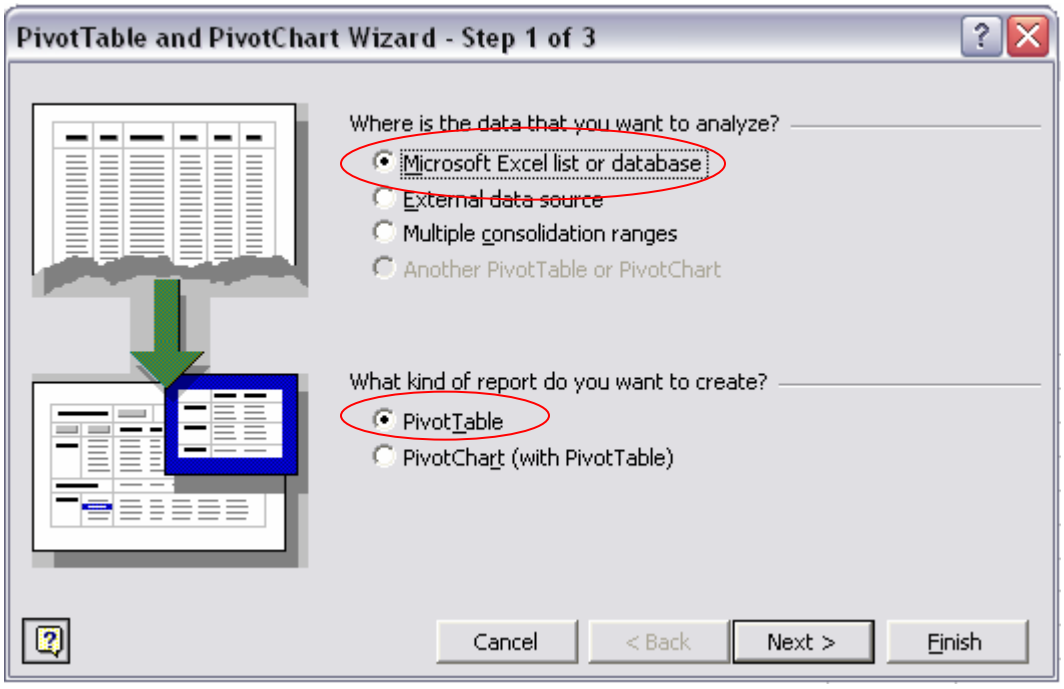
2) Arrange the data so that columns represent fields and rows represent records.

3) Start the PivotTable wizard. There are two ways to do this.
    a. Click on the Pivot Table Wizard button (  ) in the standard toolbar.

b. Go to: **View → Toolbars** and select **Pivot Table Wizard.** The pivot table toolbar will then be visible. Click on **Wizard** in the **PivotTable** pull-down menu on the toolbar.



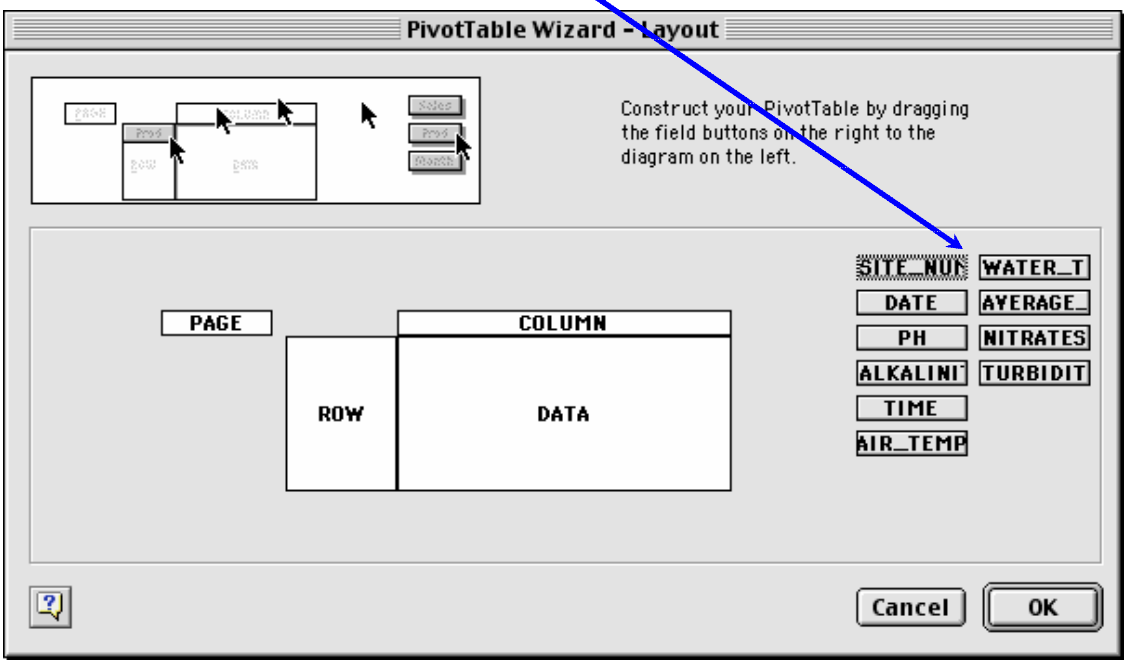

4) The first step of the pivot table wizard will then appear as a window. For this example, a pivot table will be created from an Excel database. Select the **Microsoft Excel List or Database** option and the **Pivot Table** option and click **Next.**



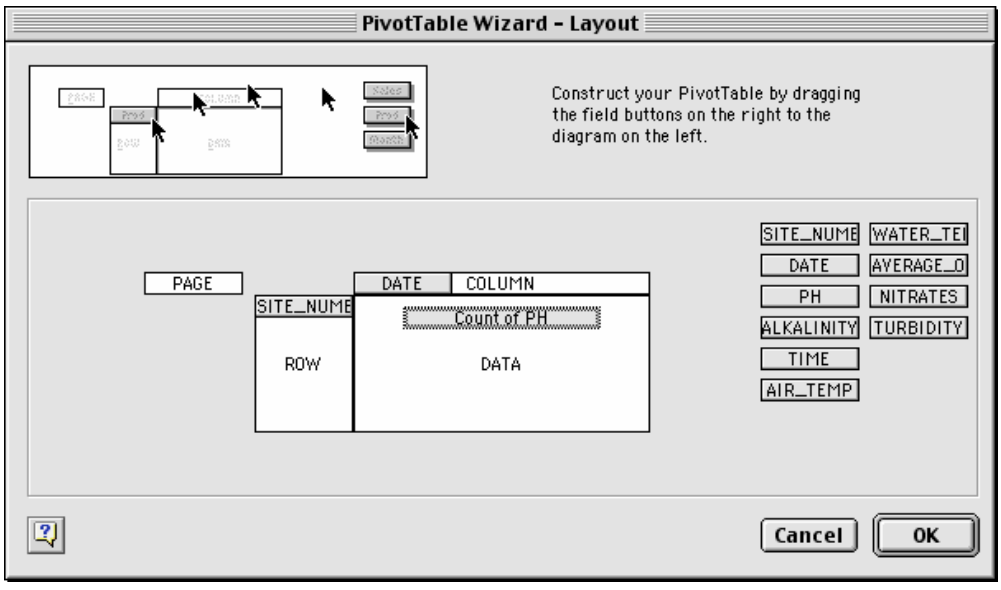

5) The next window will be **PivotTable Wizard Step 2 of 3**. Select the spreadsheet that contains the source data. In the spreadsheet, select the range of cells containing the data you'll be working with, including the column headings (a must!). Select the entire range at once. In the example window below, the "rvsdata1101!$B$2:$K$237" text in the box refers to the file name (rvsdata1101) range of cells (B$2:$K$237) that were selected. Click the **Next** button.


6) Now you'll see the final step of the pivot table wizard (**PivotTable Wizard Step 3 of 3** – see below). Click the appropriate option to tell the program whether you want the table in a new worksheet, or in the one you are working in (in this case, it will place the table in the existing worksheet with the upper left corner in cell I26. Note that you can specify a location by clicking the icon just to the right of the box and selecting the location in the spreadsheet. Click the **Layout** button.

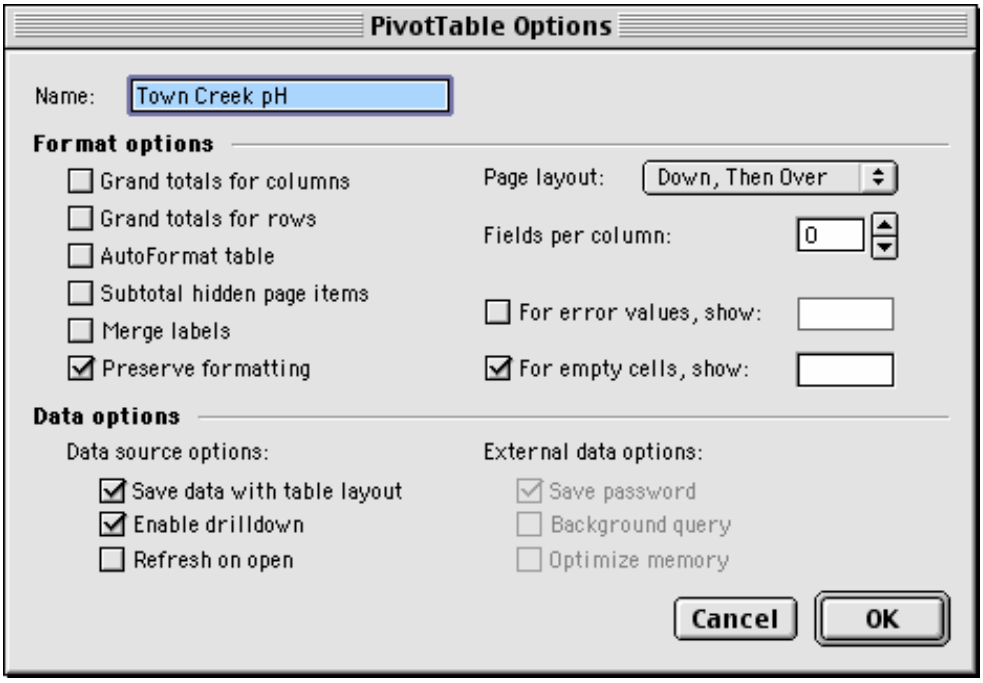
7) You'll see the following window (**PivotTable – Layout**). The boxes on the right are the column headings ("field buttons") in the cell range you selected in Step 6 (above).




8) Select and drag each of the field buttons to its appropriate place in the diagram. In this case, we want to create a table with the sites on the left of the table and the dates across the top. This is shown by the window below. Note that you can double click on the **Count of pH** field and you can proceed to the procedures described in step 13 at this point. After dragging the fields to their desired locations and/or selecting the desired summary statistics, Click OK to go back to the **PivotTable Wizard Step 3 of 3.**



9) Next, click the **Options** button and make selections so that the options window looks like the window below or make modifications to suit your needs, and then click **OK.**



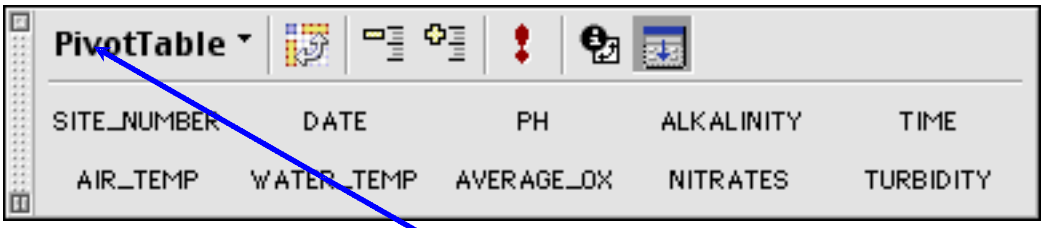

10) The window for **PivotTable Wizard Step 3 of 3** will be active again. Click **Finish** and the table will appear in the spreadsheet. Here's the upper left corner of the table based on this example. Note the field names.

| 26 | Count of PH | DATE | | | |
|---|---|---|---|---|---|
| 27 | SITE_NUMBER | 9/3/00 | 9/30/00 | 10/1/00 | 10/28/00 |
| 28 | BGR 1.28 | 0 | 0 | 0 | 0 |
| 29 | CFB&W | 0 | 1 | 0 | 1 |
| 30 | ESHC 0.30 | 0 | 0 | 0 | 0 |
| 31 | FC 1.16 | 0 | 0 | 0 | 0 |
| 32 | FC 4.40 | 0 | 0 | 1 | 0 |
| 33 | FC 4.53 | 0 | 0 | 0 | 0 |
| 34 | LOR 0.11 | 0 | 0 | 0 | 0 |
| 35 | MB 0.80 | 0 | 0 | 0 | 0 |

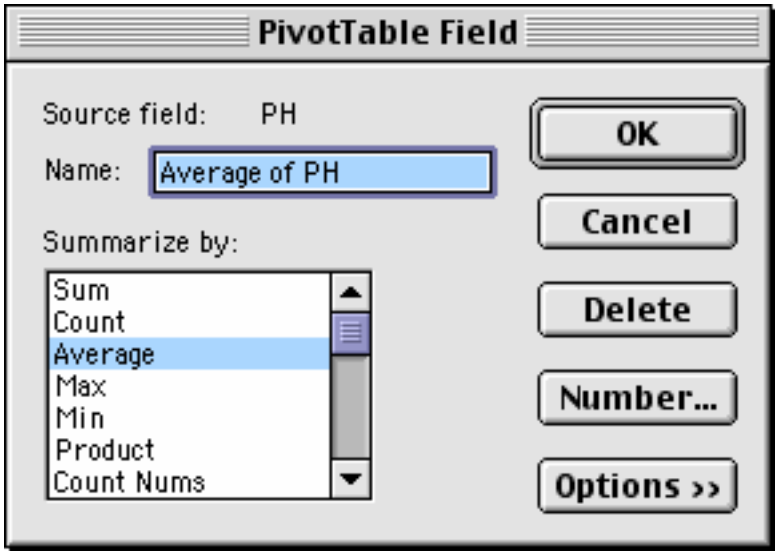
11) If the values for pH (in this case) are not the ones from the source data, it may be because they are actually calculated values. In this case, the values that appear in the cells are actually a count of the number of values in each cell of the source data. This is stated in the upper left cell which says **Count of PH**. What if we want to show the actual pH values? Unfortunately, PivotTables only display the results of calculations (functions). In this case, the table is displaying the results

of calculation which counts the number of values in each cell. This is easy to work around. If we wish to view daily results for each site, we just need to select another function that will return the original values.

12) To change the type of calculation, the Pivot Table toolbar will need to be open. If it was not opened in Step 3 of these directions, open the **View** menu by clicking on it, move your cursor to **Toolbars**, and select **PivotTable.** This toolbar will then appear:




13) Select a cell from the results area or a data label (**Count of pH**) in order to alter the type of calculation. Click on **PivotTable** in the upper left corner of the **PivotTable toolbar**. This is a pull-down menu. Select **Field Settings** from this menu. The **Field** Settings option will only be available if a cell is selected as described at the beginning of this step. The **PivotTable Field** window will open. In the example below, **Average** was selected.




14) Click **OK** to view your completed pivot table.

## 3.3 Trend Analysis

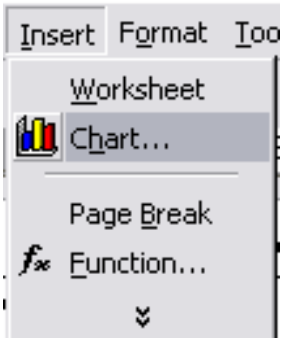
Most trend analysis that uses long term monitoring data is conducted to determine if there are changes in water quality over time. It can even be used on data that spans a relatively short period of time to show, for example, changes in water quality throughout the duration of a storm event. Trend analysis can be used to show spatial trends, like changes in water quality along the length of a stream. Whether it is applied temporally or spatially, trend analysis can be used to identify areas where water quality is being improved or degraded.

### 3.31 Graphical Trend Analysis Methods

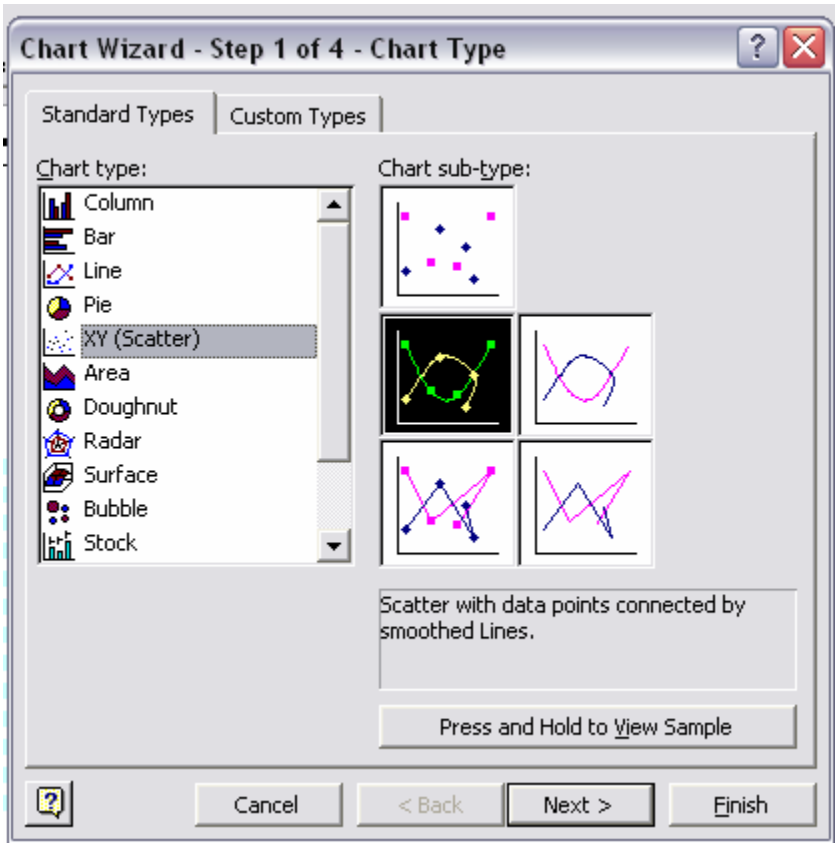
Spreadsheet programs such as Microsoft Excel are a popular method for the easy creation of graphs showing trends in data. Time series plots are created easily within this program. Due to the seasonal variability of water quality measurements, however, identifying trends can still be difficult. Software based regression analysis can be applied in order to "smooth out" the variation and show overall trends over a period of time. Regression analysis can be easily applied within Excel using a trendline. The methods below list the steps necessary for creating a simple time series plot and add a trendline to see if there is a trend in the data.

1. The quickest and easiest way to start a time series plot is to highlight the two columns (or rows) of data that you will be using. Highlight the values within the date column/row that you wish to use for the graph and, while holding the control key down, select the corresponding values for your parameter as well.

2. Now that your data is selected, there are two ways to get to the chart wizard.
   a. Click the **chart wizard** button on your tool bar. 
   b. Click on the **Insert** pull-down menu and then click on **Chart**.

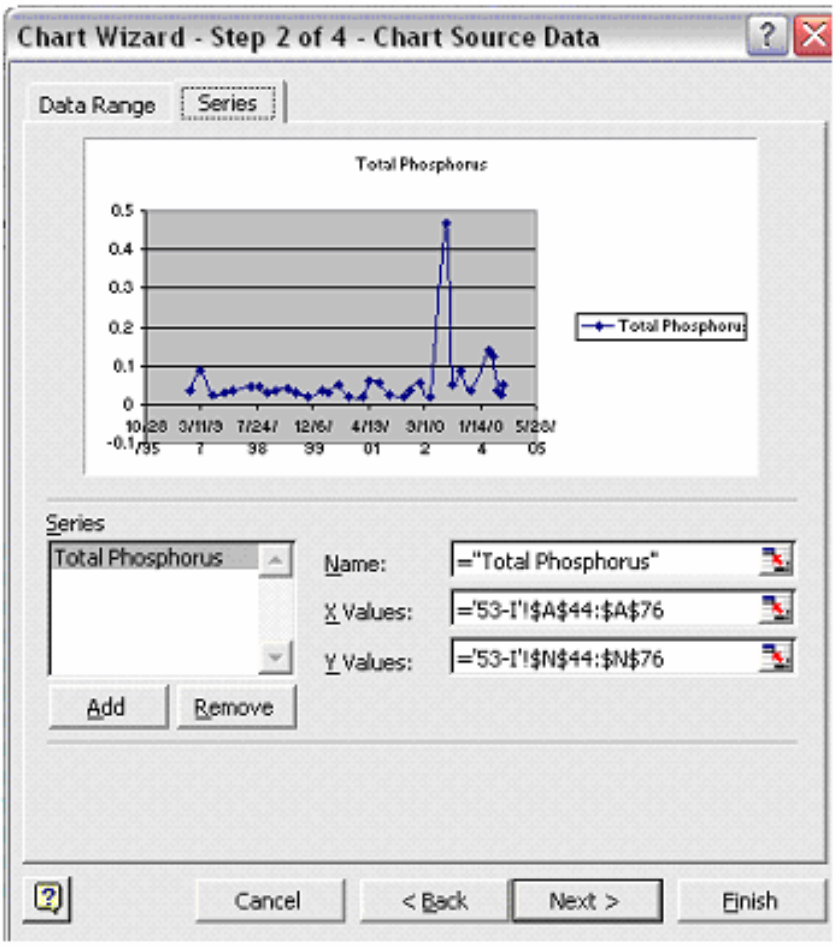



3. You are now at **Step 1 of 4** in the chart wizard process. Select **XY (Scatter)** from the list of chart types. You may choose what you want the chart to look like from the sub-type options on the right. Click **Next >** when you are finished.




4. When you get to **Step 2**, you will see a preview of your chart. Click the **Series** tab.

5. At this point, you can enter a name for your data series in the **Name** box, check to see if your graph will turn out the way you want it to. If you want to add additional data series to the chart, you can use the **Add** button to add another data series for the purpose of comparing data sets. Once everything looks the way you want it to, proceed to the next step by clicking **Next**. At any point, from this step forward, you can click the **Finish** button and skip to **Step 9** if you are satisfied with the appearance of the graph. However, going through all the steps will result in a more presentable graph.




6. In **Step 3,** you can edit details of your chart such as the chart title and axis labels. Click **next** when you are finished to go to the next step.

7. In **Step 4** of the chart wizard process, simply select where you want the chart to appear and click **finish**.

8. Your time series graph is now complete. There are several aesthetic alterations that can be made to the graph at this point by right clicking on the axis, data series, or chart area and using the respective formatting windows.

9. To apply regression to your graph to try to find a trend, right click on your data series and select **Add Trendline.**

10. The **Add Trendline** window will now be visible on your screen. Select **Linear** for the graph type, and then click on the **Options** tab. Under this tab, you may choose to display the equation on the chart, or display the r-squared value if you so desire. Press **OK.**

11. A trendline will now be visible on your chart. The slope of this line will indicate the direction of the trend in your data.

If a linear trendline doesn't show a trend, there are other types of trendlines to try. The types available in Microsoft Excel include logarithmic, polynomial, power, exponential, and moving average trendlines. A moving average trendline is particularly useful for use on long-term monitoring data sets from sites that have experienced both upward and downward trends over time.

### 3.32 Statistical Trend Detection Methods

If a trend is not easily detected by a time series plot or linear regression, this does not necessarily mean that it does not exist. There may simply be some complicating factors involved that will necessitate further statistical analysis. There are many factors that can affect the determination of trends. These include seasonal variation, day-to-day variation, and concentrations that vary with flow. One thing to consider when conducting trend analysis is to try to compare "apples to apples" instead of "apples to oranges." For example, instead of viewing all data results at once, view just the results for one season (or month) at a time to determine a trend. This concept and others are incorporated into some more technical methods of statistical analysis for the detection of trends. Some of the concepts introduced by the more technical methods found in *Statistical Methods in Water Resources* by D.R. Helsel and Hirsch's *Statistical Methods in Water Resources* and the EPA *Guidance Manual for Data Quality Assessment (G-9)* can be applied to the trend analysis that can be done with Excel. Most of the descriptions of statistical methods found in Helsel and Hirsch are very technical while the EPA guidance manual (EPA QA/G-9) and, hopefully, the manual you are reading right now do a better job of explaining these methods in a more understandable fashion.

The different methods mentioned in *Statistical Methods in Water Resources* include the Mann-Kendall test, parametric regression, LOWESS, seasonal Kendall test, data transformations, and step-trend analysis. The EPA *Guidance for Data Quality Assessment* covers trend detection methods such as regression, Sen's slope estimator, seasonal Kendall slope estimator, and hypothesis tests for detecting trends. A concept behind some types of statistical analysis for trend detection involves disproving the null hypothesis, which states that there is no trend. In other words, if there is not enough proof to say there is not a trend, than a trend may exist. Some of the tests and techniques do approximately the same thing that the Excel method described in Section 2.31 can do for you. Some involve data transformations (natural log) to improve the performance of statistical tests. Others involve techniques to determine a trend by reducing variability (seasonality) or by reducing the influence of flow on results.

LOWESS (LOcally WEighted Scatterplot Smooth) is a nonparametric method used to create a smooth line through a scatterplot. It is useful when there is a non-linear relationship between time (x) and concentration (y). Adding a moving-average trendline to a scatter plot in Microsoft Excel will essentially accomplish this type of plot.

Dealing with seasonality: There are many exogenous variables (external factors) that can affect sample results and make trend detection difficult. The variation of environmental conditions from season to season is one of these exogenous variables. Sample results vary from season to season within a year. This variation, due to weather, biological activity, natural activities (wildlife), agricultural activity, groundwater influence, and surface runoff influence, can make discerning a trend from an entire data set difficult. A particular level of discharge can either come from either ground water or surface runoff, depending on the time of the year, so seasonal stratification makes more sense than flow stratification for trend analysis (unless there is enough data to stratify by both season and flow).

In order to minimize the influence of seasons, data can be stratified by season. This way, the sample results within each data set will have been influenced by similar environmental factors. Finding a trend from summer data, for example, may be more successful than trying to find a trend from data from all seasons. There may be upward trends in some seasons and downward trends in others, even at the same monitoring site. Trends may appear in seasonally stratified data that do not appear in the entire data set. This may happen if both upward and downward trends exist for separate seasons that may cancel each other out when all the data is combined. Seasonal strata can be quarterly (four per year) or monthly (twelve per year). Quarterly stratification will yield a more manageable amount of results than monthly stratification. Once data has been stratified, the Excel method described in this document can be applied to each season's data set to create time series plots. The seasonal Kendall test and regression analysis are two statistical methods that can be applied to seasonally stratified data in order to find a trend.

Sen's Slope Estimator: For this nonparametric alternative method for finding a slope, the slopes between each set of points in time are calculated first. The median of all these slopes is then used as the overall slope.

Seasonal Kendall Test: This slope test can be used to account for cyclical trends. The concept presented by this test is that a trend may be evident if slope is calculated for each season, month, or week.

Mann-Kendall Trend Test: This method is used for testing a hypothesis for the purpose of trend detection. This test involves calculating the statistic S by examining the individual slopes between all possible pairs of data. A large negative value for S indicates a decreasing trend. A large positive S value represents an increasing trend. The null hypothesis, or $H_0$, is that there is no trend. The alternative hypothesis, $H_A$, is that there is either an upward trend or a downward trend.

To calculate the Mann-Kendall trend test, list all observations in chronological order from left to right horizontally across the top of the table beginning in the same corner of a table as the horizontal lists. Also list all measurements except for the last chronologically vertically from the top to bottom along the left side of the table. Each measurement is then compared to previous measurements to determine whether there is a positive difference or a negative difference.

Within this matrix, the horizontal measurements are compared with those of vertical measurements. The value from the vertical axis is subtracted from the value of the each measurement on the horizontal axis. A plus or a minus is recorded to indicate whether the relationship is positive or negative (values of 0 are not recorded on the table). The number of pluses and the number of minuses are then added for each row and totaled at the bottom of the table. The total number of minuses is subtracted from the total number of pluses.

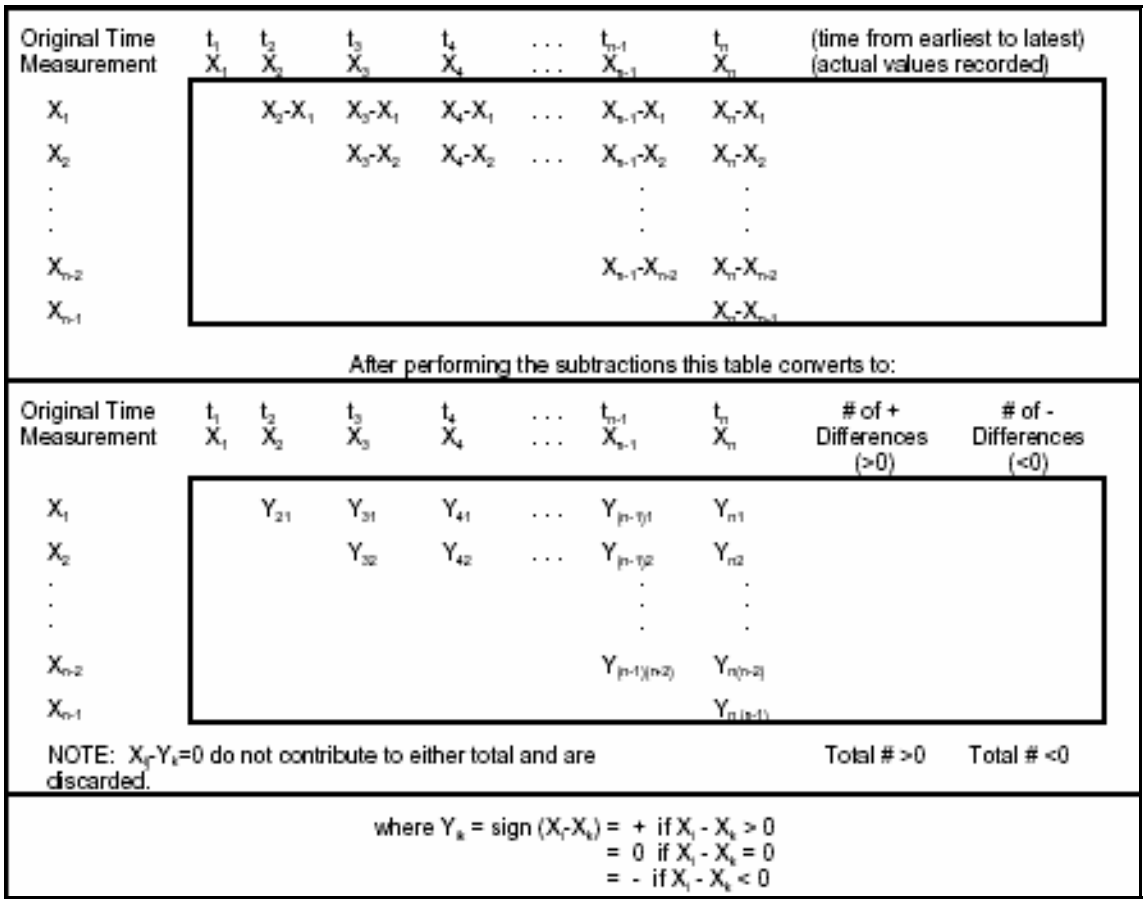



**Figure 17. "Upper Triangular" Data for Basic Mann-Kendall Trend Test with a Single Measurement at Each Time Point (EPA *Guidance for Data Quality Assessment*)**

Consider 5 measurements ordered by the time of their collection: 5, 6, 11, 8, and 10. This data will be used to test the null hypothesis, $H_0$: no trend, versus the alternative hypothesis $H_1$ of an upward trend at an $\alpha = 0.05$ significance level.

STEP 1: The data listed in order by time are: 5, 6, 11, 8, 10.

STEP 2: A triangular table (see Box 4-6) was used to construct the possible differences. The sum of signs of the differences across the rows are shown in the columns 7 and 8.

| Time | 1 | 2 | 3 | 4 | 5 | No. of + | No. of - |
|---|---|---|---|---|---|---|---|
| Data | 5 | 6 | 11 | 8 | 10 | Signs | Signs |
| 5 | | + | + | + | + | 4 | 0 |
| 6 | | | + | + | + | 3 | 0 |
| 11 | | | | - | - | 0 | 2 |
| 8 | | | | | + | 1 | 0 |
| | | | | | | 8 | 2 |

STEP 3: Using the table above, S = 8 - 2 = 6.

STEP 4: From Table A-11 of Appendix A for n = 5 and S = 6, p = 0.117.

STEP 5: Since S > 0 but p = 0.117 ≮ 0.05, the null hypothesis is not rejected. Therefore, there is not enough evidence to conclude that there is an increasing trend in the data.

**Figure 18. An Example of Mann-Kendall Trend Test for Small Sample Sizes (EPA *Guidance for Data Quality Assessment*).**

To save a little time, an equation can be used to arrive at the final table in Microsoft Excel. An if-then equation like =IF(H15<0,"-","+") can be used. This equation will determine whether or not the value in a cell is below zero and if it is, it will display a negative sign in its cell. It will display a positive sign for every value greater than or equal to zero.

Create a copy of the table containing the difference calculations and replace the values in the copy with the if-then equation. Start by placing the equation in one of the cells and making sure that it works properly. Make sure the cell reference (H15 in the example) points to the corresponding place in the original table. Copy the equation to the other cells within the table where it is needed. If the cell reference is correct in the first cell, it should be correct in the others as well because the cell reference within the equation based upon the receiving cells position relative to the cell the equation is copied from. Zero values will have to be entered manually if an if-then equation if an if-then equation such as the example is used because zero values will be transformed into + signs when the equation is initially copied across the table.

Table 3. Table A-11 from Appendix A of the EPA *Guidance for Data Quality Assessment*.

| | n | | | | | n | | |
|---|---|---|---|---|---|---|---|---|
| S | 4 | 5 | 8 | 9 | S | 6 | 7 | 10 |
| 0 | 0.625 | 0.592 | 0.548 | 0.540 | 1 | 0.500 | 0.500 | 0.500 |
| 2 | 0.375 | 0.408 | 0.452 | 0.460 | 3 | 0.360 | 0.386 | 0.431 |
| 4 | 0.167 | 0.242 | 0.360 | 0.381 | 5 | 0.235 | 0.281 | 0.364 |
| 6 | 0.042 | 0.117 | 0.274 | 0.306 | 7 | 0.136 | 0.191 | 0.300 |
| 8 | | 0.042 | 0.199 | 0.238 | 9 | 0.068 | 0.119 | 0.242 |
| 10 | | 0.0083 | 0.138 | 0.179 | 11 | 0.028 | 0.068 | 0.190 |
| 12 | | | 0.089 | 0.130 | 13 | 0.0083 | 0.035 | 0.146 |
| 14 | | | 0.054 | 0.090 | 15 | 0.0014 | 0.015 | 0.108 |
| 16 | | | 0.031 | 0.060 | 17 | | 0.0054 | 0.078 |
| 18 | | | 0.016 | 0.038 | 19 | | 0.0014 | 0.054 |
| 20 | | | 0.0071 | 0.022 | 21 | | 0.00020 | 0.036 |
| 22 | | | 0.0028 | 0.012 | 23 | | | 0.023 |
| 24 | | | 0.00087 | 0.0063 | 25 | | | 0.014 |
| 26 | | | 0.00019 | 0.0029 | 27 | | | 0.0083 |
| 28 | | | 0.000025 | 0.0012 | 29 | | | 0.0046 |
| 30 | | | | 0.00043 | 31 | | | 0.0023 |
| 32 | | | | 0.00012 | 33 | | | 0.0011 |
| 34 | | | | 0.000025 | 35 | | | 0.00047 |
| 36 | | | | 0.0000028 | 37 | | | 0.00018 |
| | | | | | 39 | | | 0.000058 |
| | | | | | 41 | | | 0.000015 |
| | | | | | 43 | | | 0.0000028 |
| | | | | | 45 | | | 0.00000028 |

**Box 4-9: Directions for the Mann-Kendall Procedure Using Normal Approximation**

If the sample size is 10 or more, a normal approximation to the Mann-Kendall procedure may be used.

STEP 1: Complete steps 1, 2, and 3 of Box 4-7.

STEP 2: Calculate the variance of S: $V(S) = \frac{n(n-1)(2n+5)}{18}$.

If ties occur, let g represent the number of tied groups and $w_p$ represent the number of data points in the $p^{th}$ group. The variance of S is: $V(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{g} w_p(w_p-1)(2w_p+5)\right]$

STEP 4: Calculate $Z = \frac{S-1}{[V(S)]^{1/2}}$ if S > 0, Z = 0 if S = 0, or $Z = \frac{S+1}{[V(S)]^{1/2}}$ if S < 0.

STEP 5: Use Table A-1 of Appendix A to find the critical value $z_{1-\alpha}$ such that 100(1-α)% of the normal distribution is below $z_{1-\alpha}$. For example, if α=0.05 then $z_{1-\alpha}$=1.645.

STEP 6: For testing the hypothesis, $H_0$ (no trend) against 1) $H_1$ (an upward trend) – reject $H_0$ if $Z > z_{1-\alpha}$, or 2) $H_2$ (a downward trend) – reject $H_0$ if Z < 0 and the absolute value of $Z > z_{1-\alpha}$.

**Figure 19. Directions for the Mann-Kendall Procedure Using Normal Approximation - for Samples Sizes Greater Than 10 (from EPA *Guidance for Data Quality Assessment).***

A test for an upward trend with α=.05 will be based on the 11 weekly measurements shown below.

STEP 1: Using Box 4-6, a triangular table was constructed of the possible differences. A zero has been used if the difference is zero, a "+" sign if the difference is positive, and a "-" sign if the difference is negative.

| Week | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | No. of | No. of |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | 10 | 10 | 10 | 5 | 10 | 20 | 18 | 17 | 15 | 24 | 15 | + Signs | - Signs |
| 10 | | 0 | 0 | - | 0 | + | + | + | + | + | + | 6 | 1 |
| 10 | | | 0 | - | 0 | + | + | + | + | + | + | 6 | 1 |
| 10 | | | | - | 0 | + | + | + | + | + | + | 6 | 1 |
| 5 | | | | | + | + | + | + | + | + | + | 7 | 0 |
| 10 | | | | | | + | + | + | + | + | + | 6 | 0 |
| 20 | | | | | | | - | - | - | + | - | 1 | 4 |
| 18 | | | | | | | | - | - | + | - | 1 | 3 |
| 17 | | | | | | | | | - | + | - | 1 | 2 |
| 15 | | | | | | | | | | + | 0 | 1 | 0 |
| 24 | | | | | | | | | | | - | 0 | 1 |
| | | | | | | | | | | | | 35 | 13 |

STEP 2: S = (sum of + signs) - (sum of - signs) = 35 - 13 = 22

STEP 3: There are several observations tied at 10 and 15. Thus, the formula for tied values will be used. In this formula, g=2, $t_1$=4 for tied values of 10, and $t_2$=2 for tied values of 15.

$$V(S) = \frac{1}{18}\left[11(11-1)(2(11)+5) - [4(4-1)(2(4)+5) + 2(2-1)(2(2)+5)]\right] = 155.33$$

STEP 4: Since S is positive: $Z = \frac{S-1}{[V(S)]^{1/2}} = \frac{22-1}{(155.33)^{1/2}} = \frac{20}{12.46} = 1.605$

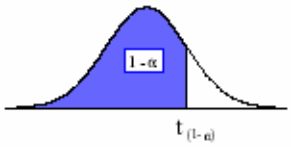
STEP 5: From Table A-1 of Appendix A, $z_{1-.05}$=1.645.

STEP 6: $H_1$ is the alternative of interest. Therefore, since 1.605 is not greater than 1.645, $H_0$ is not rejected. Therefore, there is not enough evidence to determine that there is an upward trend.

**Figure 20. Example of Mann-Kendall Trend Test by Normal Approximation for Sample Sizes of 10 or More (From EPA Guidance for Data Quality Assessment).**

**Table 4. Critical Values of t Distribution. Table A-1 from Appendix A of the EPA Guidance for Data Quality Assessment (for Steps 5-6 in Figure 20).**

TABLE A-1: CRITICAL VALUES OF STUDENT'S t DISTRIBUTION




| Degrees of Freedom | 1 - α | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | .70 | .75 | .80 | .85 | .90 | .95 | .975 | .99 | .995 |
| 1 | 0.727 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 0.617 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 0.584 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 0.569 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 0.559 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 0.553 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 0.549 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 0.546 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 0.543 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 0.542 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 0.540 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 0.539 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 0.538 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 0.537 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 0.536 | 0.691 | 0.866 | 1.074 | 1.34 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 0.535 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 0.534 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 0.534 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 0.533 | 0.6880 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 0.533 | .687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 0.532 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 0.532 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 0.532 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | 0.531 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 0.531 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 0.531 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 0.531 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 0.530 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 0.530 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30 | 0.530 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40 | 0.529 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| 60 | 0.527 | 0.679 | 0.848 | 1.046 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| 120 | 0.526 | 0.677 | 0.845 | 1.041 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| | 0.524 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

Note: The last row of the table ( degrees of freedom) gives the critical values for a standard normal distribution (z), e.g., $t_{,0.95} = z_{0.95} = 1.645$.

Alternative Methods and Data Transformations: Some data sets may have non-linear trends that won't be found using methods for determining a linear slope coefficient. In these cases (although not in all cases) transforming data before trend analysis may increase the chance of success in finding a linear trend. Transforming the data into natural log units is one way to do this. Create a linear trend line using the transformed

data by using the methods described in a text book for linear regression or by using Microsoft Excel (the easier way) to create a trendline through a time series plot or a time graph of the data. Once a trend is found, the trend slope will be expressed in log units and the percentage of change can be calculated by using the equation: $(e^m - 1)*100$, where m is the slope of the linear trend in log units. Remember that m = slope in the equation of a line (y = mx + b). Therefore, in the equation y = 2x + 3, the slope is equal to 2. For example, the slope of the linear trend of the natural logs of spring total suspended solids results from the Clearwater River at the USGS gauge near the town of Plummer, Minnesota is .1804. When m (in log units) =.1804, the percentage of increase in spring total suspended solids concentrations each year is 19.77%.

If events have occurred within the watershed of a particular monitoring site that may have had an effect on water quality and the dates of these actions are known, they should be considered during trend analysis. These actions could include the removal of a dam, an upgrade to a waste water treatment plant, erosion control projects, impoundments, implementation of buffers within the watershed, and lake restoration projects. The original data set may be split into "before" and "after" data sets. Make sure that the data split is based on the timing of the event and not based upon an examination of water quality data, or bias may be introduced into the analysis processes and trend analysis may show changes that aren't really there. For more information on statistical methods for trend detection, consult a statistics textbook or a free resource like the EPA *Guidance for Data Quality Assessment – Practical Methods for Data Analysis – EPA QA/G-9 – QA00 Update* that is available for free online at http://www.epa.gov/quality/qs-docs/g9-final.pdf.

### 3.4 Developing Load Estimates

Load estimates are used to determine the mass of a substance being carried by a river or stream through a sampling site within a particular amount of time. Loads can be calculated on an annual or a seasonal basis, depending upon how much data is available. Annual loads can only be accurately estimated when there is a full year's worth of data. If a full year's worth of data is not available, seasonal estimates can be done for the period of time for which there is data available (i.e. April through October). By comparing annual or seasonal loads, the relative impact that a watersheds or subwatershed is having on water quality can be quantified.

There are a number of software programs that calculate loads and can estimate annual/seasonal loads based upon flow and water quality data. Some of these are free such as FLUX and Basins. Others can be somewhat expensive. The free versions are, in some cases, preferred by resource professionals because the models and the methods used within the models do not change as much as purchased software. This makes it easier to compare results from different monitoring programs. The RLWD currently uses FLUX for load estimation. It is a DOS-based program distributed by the U.S. Army Corps of Engineers that was developed by Dr. William W. Walker. Some of the advantages of this program are that it is reliable, relatively accurate, and provides a lot of information for each data set. There are some negatives and annoyances that have been encountered with

FLUX. The old version that was not Y2K compliant (this has since been fixed). Data has to be transferred into Lotus spreadsheets before it can be loaded into the model - causing extra work for Excel and Access users. The program is very fussy about the organization of data within the spreadsheets. The user manual does not always cover the quirks of the program very well. This section will provide some tips that will hopefully make the learning process a little smoother for those who wish to use the FLUX program.

The first step in creating load estimates is the collection of data. Higher numbers of samples will generally result in load estimations of higher accuracy. Also, the collection of flow data is very important. Daily average flow data should be obtained for the entire period of record that will be modeled. This can be done using flow data from a nearby USGS gauge or by installing continuous stage recording equipment, collecting a range of flow measurements, and creating rating curves to estimate flows based on the stage data. For more information on stream gauging, flow monitoring, and the creation of rating curves, see Section 9.0 of the Standard Operating Procedures for Water Quality Monitoring in the Red River Watershed and Section 3.56 of this manual.

The next step is the preparation of data so that it can be used by FLUX. For this step, data can be prepared and organized in Excel much more quickly and easily than in Lotus 1-2-3. A separate work sheet is needed for each parameter and for flow. Creating a workbook for each site and worksheets for each parameter within each workbook is recommended. This is because there usually is less sampling data than flow data available.

If there is not a sample result for each day that there is a value for flow, there will be gaps in the parameter data if it is placed in a column next to the flow data (within the same table). FLUX reads from the top down in each column of data and when it encounters a blank or zero value, it stops reading values, so if there are blank cells between results, not all of the data will read by the FLUX program.

In the spreadsheet, a title on the first line of the table, and column headings in the second row are another necessity. The DATE column headings should be typed in all capital letters. Use consistent column headings for flow and other parameters. You will need to remember what these column headings are (writing them down helps) when you are telling FLUX where to find the data. Each individual worksheet within the workbook will need to be saved as a .WK1 file if it has been created in Excel.

When saving the worksheets, put them in a location where the file path is easy to remember (C:\model\Data\) The Lotus spreadsheet below is formatted to work with FLUX. Keeping track (recording) file names, column headings, and date ranges is highly recommended so a quick reference is available when bringing data into FLUX.

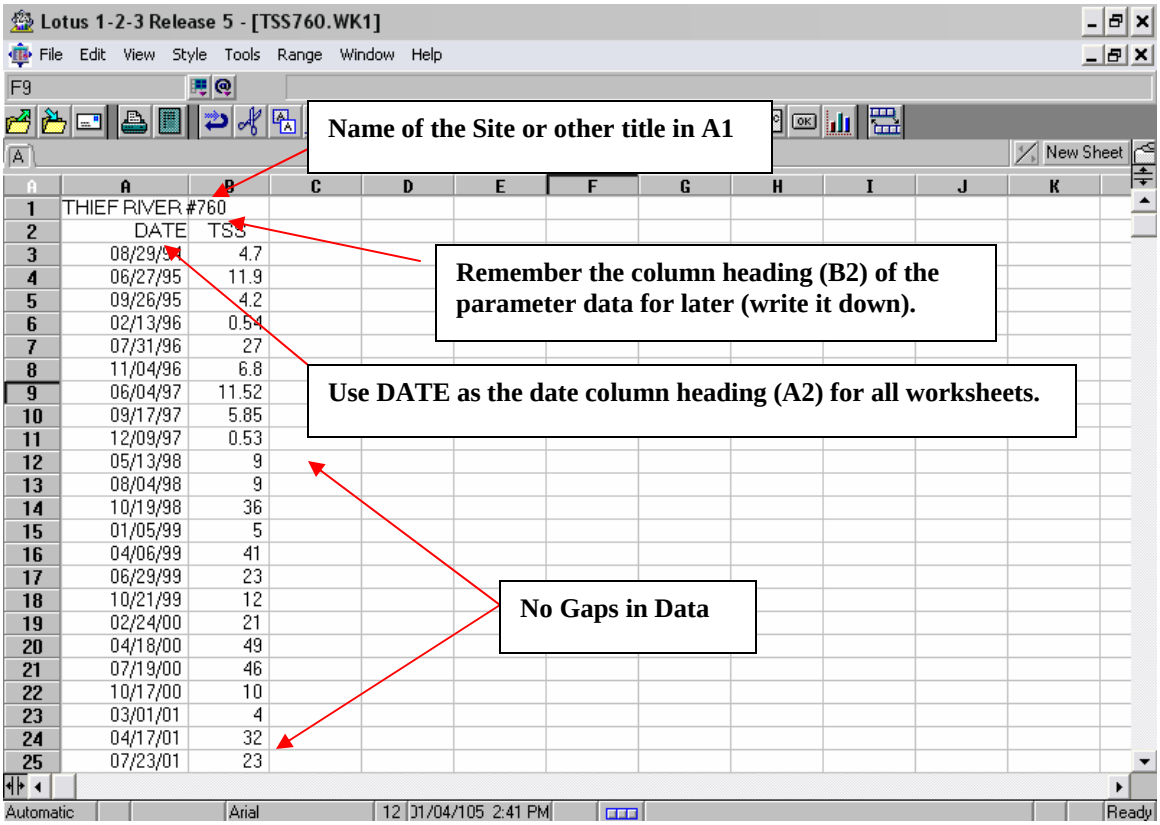



**Figure 21. Lotus Spreadsheet Configured for FLUX.**

When the data to be analyzed has been set up in this fashion, the FLUX program can be started. Once you have gotten to the main menu, you will need to tell the program to read your data. Use the arrow keys to navigate the menu system from DATA down to READ, and then down to RESET and then hit enter. The program will then switch to the FLUX INPUT SCREEN shown below.

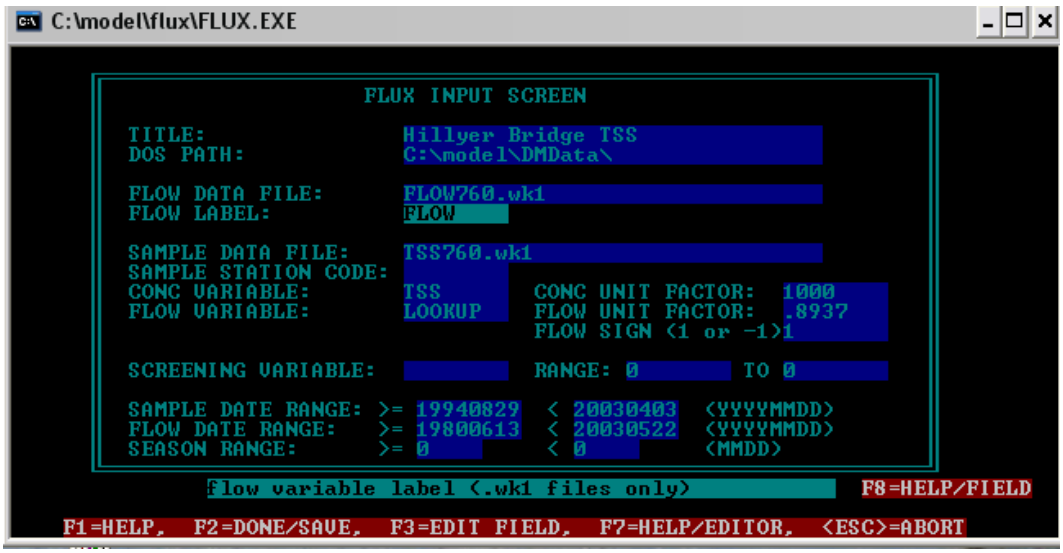



**Figure 22. FLUX Input Screen.**

On the FLUX Input Screen (Figure 12), enter a title, such as the site name and the parameter being analyzed. Then enter the **DOS PATH**, which is the location of the folder in which the data is stored (store the flow data and parameter data in the same folder) as it would appear in the address bar of Windows Explorer. In figure 12, the DOS path is C:\model\DMData\. Be sure to include a backslash ( \ ) at the end of the DOS path. If you are unsure of the exact path for your data file, use Windows Explorer to find the file and use the path shown in the address bar to get the correct file location. The **FLOW DATA FILE** is the name (FLOW760) and extension (.wk1) of the Lotus file in which the flow data is stored. **FLOW LABEL** is simply the column heading for flow within this spreadsheet. **SAMPLE DATA FILE** is the name and extension of the file containing the sample data and **CONC VARIABLE** is the column heading for the sample data. Entering **LOOKUP** for the flow variable will tell the program to lookup the corresponding flow for each sample result. **SAMPLE DATE RANGE** and **FLOW DATE RANGE** are filled in with the beginning date on the left and the ending date on the right. Press F2 when you are done. If everything goes right, you will get a screen with the information listed below that lists statistics, such as the number of flow records and the number of samples, like the one below. You can then hit escape until you get back to the main menu.

```
Locating Sample File....
OPENING SAMPLE FILE = TSS760.WK1
SAMPLE CONCENTRATION FIELD = TSS
CONCENTRATION UNITS FACTOR =     1000.000000
Flow Scale Factor =       .8937
Conc Scale Factor =   1000.0000
Reading Samples...
THIEF RIVER #760
NUMBER OF SAMPLES =         16
Reading Flows...
OPENING FLOW FILE = FLOW760.WK1
FLOW FIELD = FLOW
THIEF RIVER #760
NUMBER OF FLOW RECORDS =       6999
Substituting Daily Flows for Sample Flows
 Flow/Concentration Pairs =         16
 Missing or Zero Flows on Sample Dates =          0
```

If you receive an error instead of a list similar to the one above, you will need to check the information entered into the **FLUX INPUT SCREEN**, especially the data file location and file name. Check to make sure that the data in the spreadsheets is entered correctly, and make sure the data is arranged correctly on the spreadsheet.

Once data is loaded into FLUX, one of the programs primary functions is calculating the load over the time period specified. If multiple years of data are used, it will calculate the average annual load. If the data is stratified by season and includes multiple years of data, it can calculate the average load for each season.

One of the most time consuming parts of using FLUX is the determination of which calculation and stratification methods produce the most accurate results. The best calculation method is found first, and then that calculation method is applied to several different stratification schemes in an effort to find the lowest coefficient of variance. The coefficient of variance is a measure of the accuracy of the estimate. A lower CV means a higher level of accuracy in the model's calculations.

FLUX uses several different calculation methods:

1. Direct mean loading
2. Flow-weighted concentration (ratio estimate)
3. Modified ratio estimate
4. Regression, first order
5. Regression, second order
6. Regression, applied to individual daily flows

Fortunately, knowledge of how all these calculation methods work is not needed in order to run the model. In order to choose the best calculation method for your data, you will need to determine which method is the most accurate, or which method has the lowest coefficient of variance. FLUX calculates this value. To find the method with the lowest coefficient of variance, use your arrow keys to highlight **CALCULATE** in the main menu. Highlight **LOADS** in the submenu that appears and press the **ENTER** key. In the resulting window, there will be a list of annual load results for each calculation method. Make note of which method has the lowest CV and press the Esc key to get back to the main menu. In the example below, method 5 (CV = .147) will be the most accurate of the six methods.

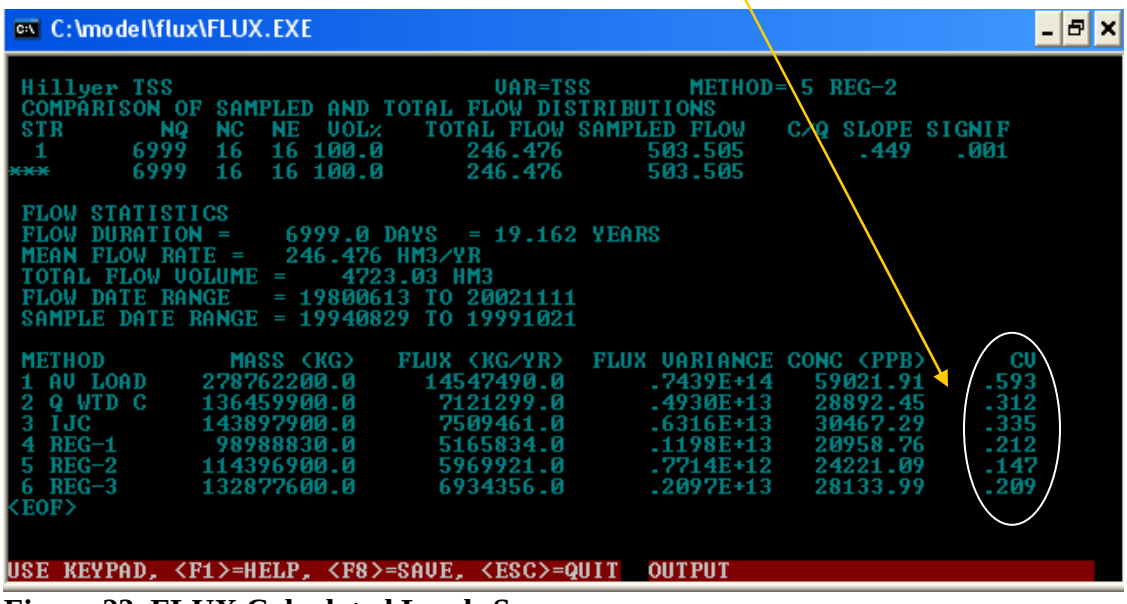



**Figure 23. FLUX Calculated Loads Screen.**

The program must now be told to use the desired calculation method (#5 in the example) for use in the subsequent calculation of loads. The program will apply the selected calculation method to each stratification method that you apply to the data. To select a calculation method, return to the **MAIN MENU** and highlight **METHOD**. Then, in the submenu, **highlight the chosen method** and press the **ENTER** key.

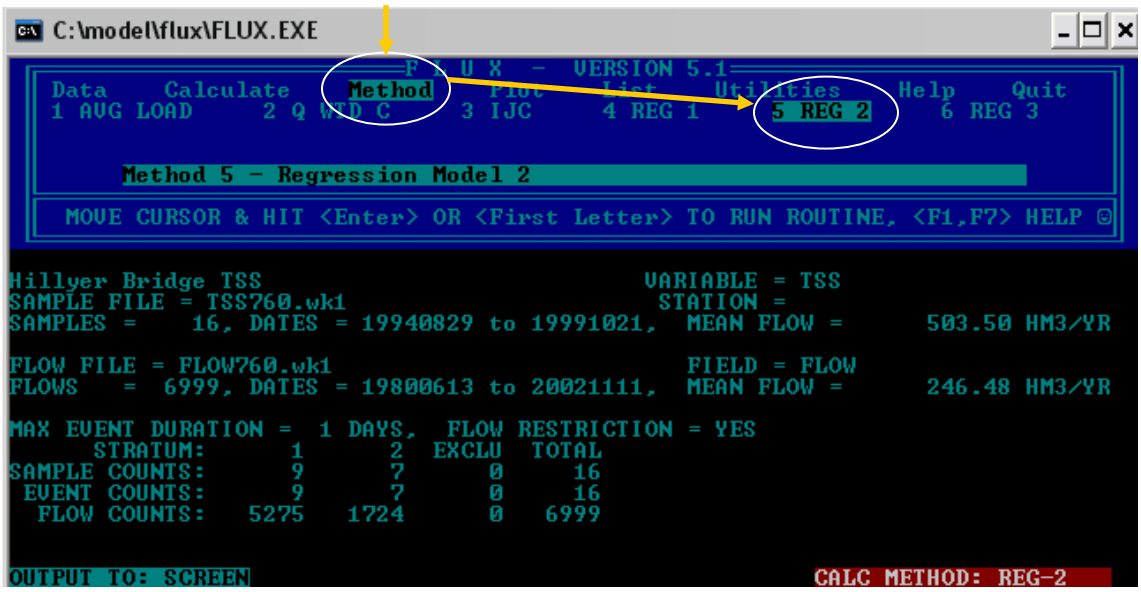



**Figure 24. Choosing a Load Calculation Method in FLUX.**

Stratification is a process that splits the data into groups by flow or by time. A maximum of five strata can be created in FLUX. Stratifying data can improve the accuracy of load estimates, as long as there are enough samples. As with finding the best calculation method, finding the best stratification method also involves trying to get the lowest CV possible. In this example, data will be stratified by flow. FLUX will automatically set the boundaries of flow strata.

To stratify using dates or another stratification system, use the General stratification option and the number of strata needed under **Stratify** in the **MAIN MENU**. Then input dates or other values to specify boundaries between strata.

After choosing the best calculation/estimation method, highlight **LIST** and then **BREAKDOWNS** in the submenu and press **ENTER** to get the breakdowns by stratum. Since the default stratification scheme is one stratum (no stratification), the breakdown results will be for one stratum the first time you do this. Breakdowns show the number of samples, flow volume per year in $HM^3$/yr, FLUX in Kg/yr, total volume in $HM^3$, total mass in Kg, mean concentration in ppb, and the coefficient of variance (CV). Note the CV value (.147 in the example below) and press **ESC** to return to the main menu. Now, you will try to use additional strata in an attempt to decrease the CV.

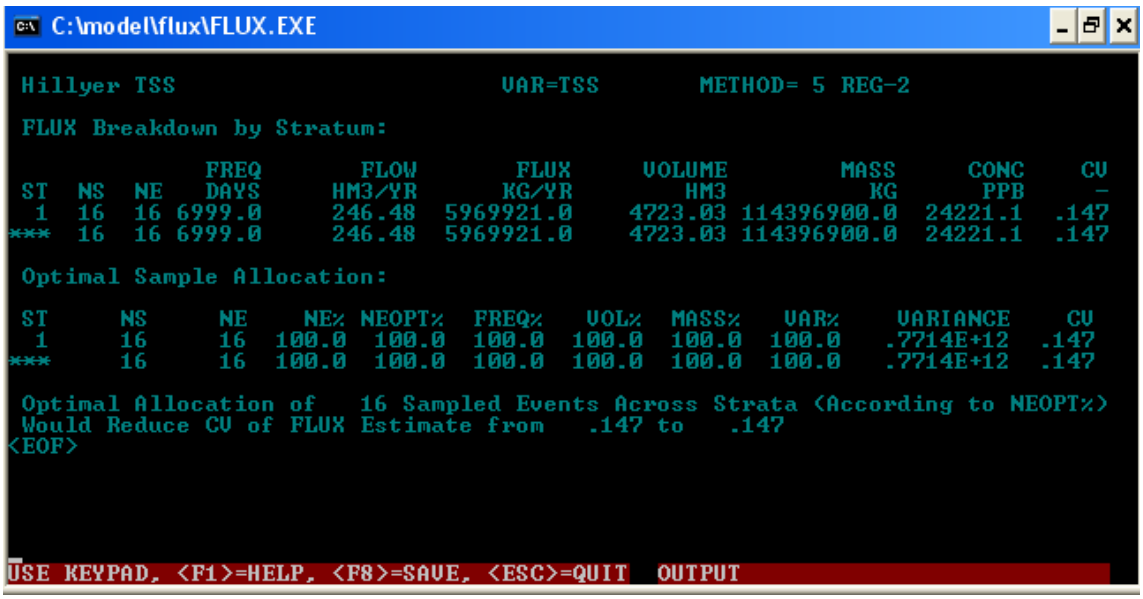



**Figure 25. Breakdowns Screen.**

This step demonstrated in the screen shot below is used to test other stratification schemes based upon flow by increasing the number of strata. In the main menu, with **DATA** highlighted, use the arrow keys to get to **Stratify**, then **Flow**, then **2 Strata**, and then press **ENTER**.

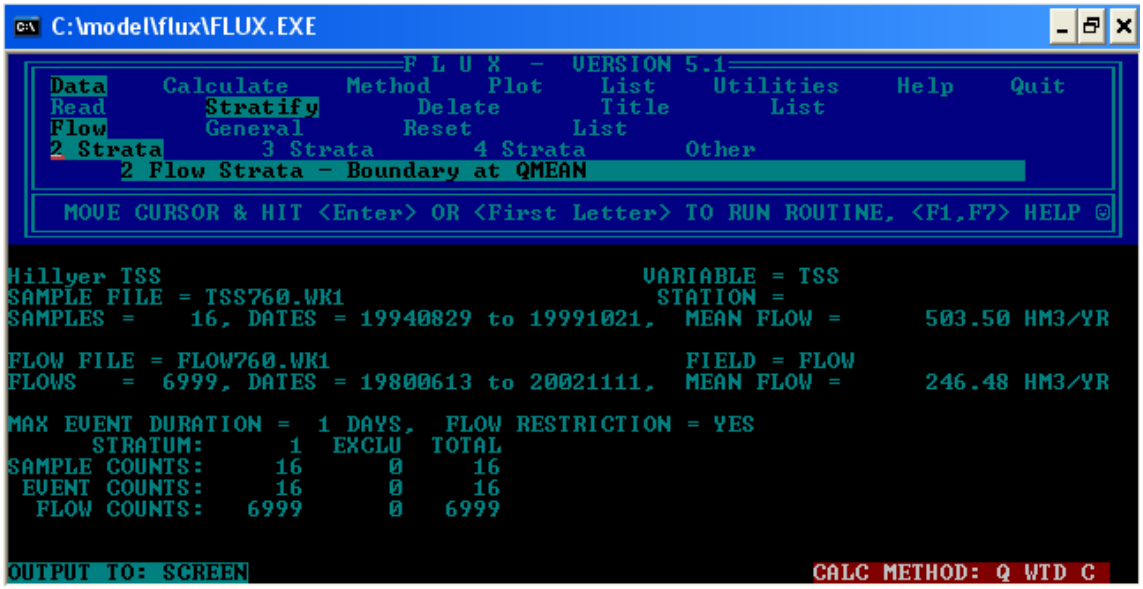



**Figure 26. Path Through the Menu to Stratification.**

FLUX will automatically stratify the flow into two categories.

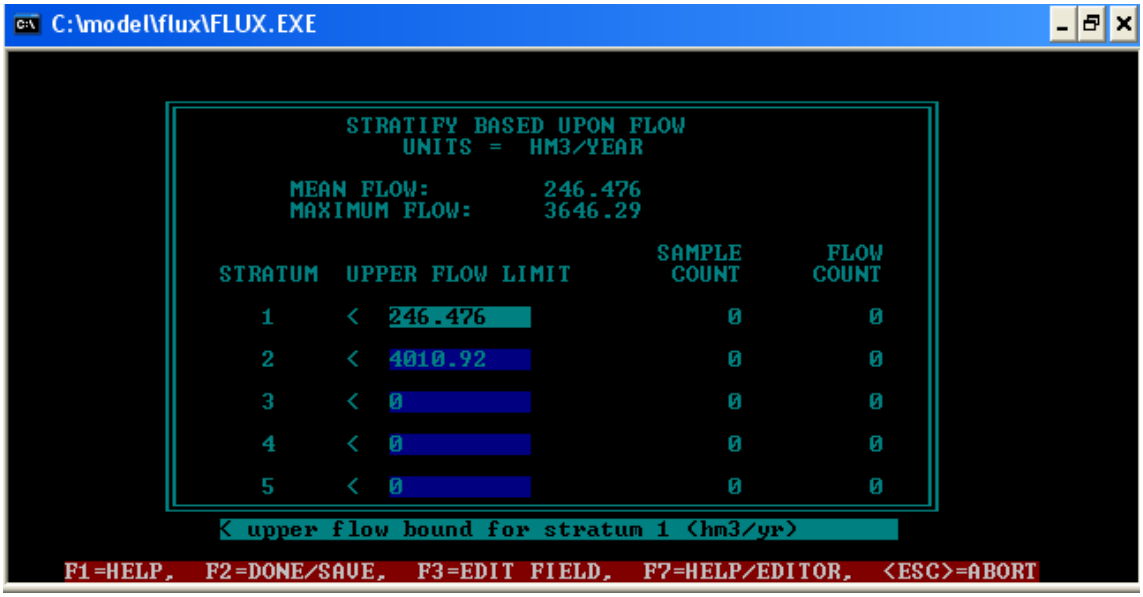



**Figure 27. Stratification Screen.**

FLUX will automatically use the mean flow volume as the boundary between Stratum 1 and Stratum 2. The flow levels for each category can be modified, but for the sake of sticking to the basics, press the **F2** key and then the **ESC** key to go back to the **MAIN MENU.** To see if this stratification improves the CV, go to **LIST** => **BREAKDOWNS** once again to see if the overall CV (the average values are in the bottom rows of each table shown on this screen) is larger or smaller than the previous CV. If it is smaller, try using 3 strata, or even 4 to find the smallest possible CV. A limiting factor for the amount of stratification that can be applied to data is the number of samples in each stratum. When there are too few samples per strata (too many strata), the FLUX program will inform the user of this problem by displaying an illegal stratification error.

After you have achieved the lowest possible coefficient of variance, record the breakdowns for the calculation method and stratification method combination with the lowest CV. Note that, in Figure 17, the CV was lowest using calculation method #5 (regression, second order) and, using this calculation method, the CV was lower with two strata than for one stratum (.130 vs. .147, respectively). Adding another strata did not reduce the CV any further.

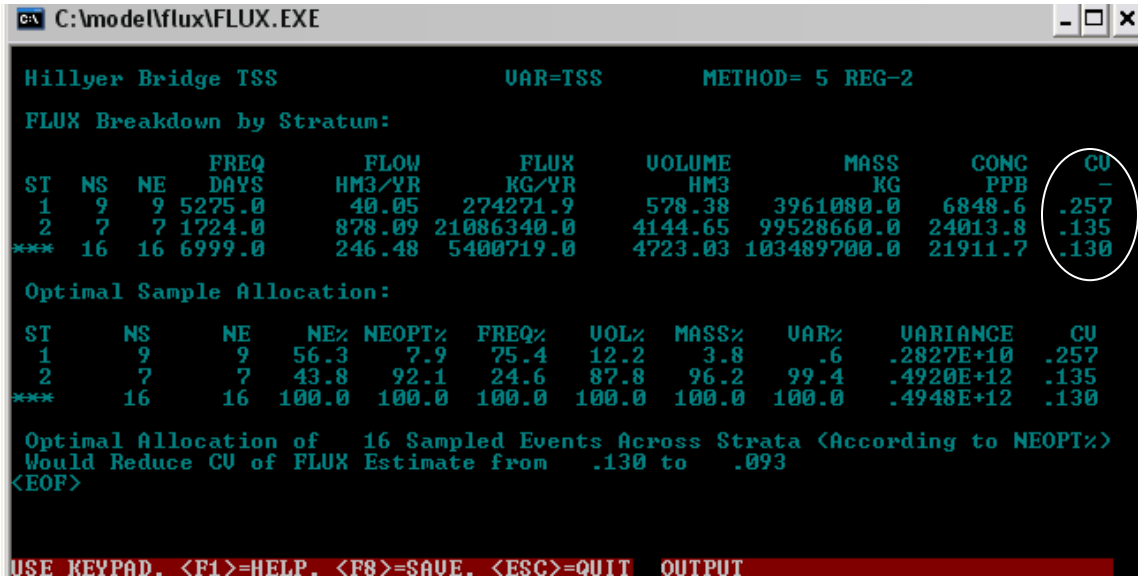



**Figure 28. Noting the Coefficient of Variance.**

When the most accurate method has been found, the values for flow (cubic hectometers per year), flux (Kilograms per year), total volume (hectometers), mass (Kilograms), and flow-weighted concentration (parts per billion) can be recorded from the breakdowns page.

FLUX can also be used to evaluate your monitoring program. Modeling results can be biased based upon the distribution of samples. Since most of the sediment and nutrient loading from rivers occurs during high flows, the majority of samples should be collected during high flows to achieve the most accurate annual load estimations. FLUX contains a function that determines the optimal percentage of samples that should be collected for each stratum. When the data has been stratified by FLUX, whether by flow or temporally, the distribution of the sample data with the optimal distribution of samples can be compared. For example, under a flow stratification system of high versus low flows, the majority of samples may have been collected during low flows, but the optimal distribution that FLUX calculates will show that the majority of the samples should be collected during high flow periods. Using this comparison, a monitoring program can be adjusted to, for example, collect more samples during high flow periods than during low flow periods if one of the goals of the program is the calculation of annual or seasonal loads. These calculations of optimal sample distributions are found in the optimal sample allocation section of the breakdowns screen (see Figure 28). **NE%** is the actual percentage of samples in each stratum. **NEOPT%** is the optimal percentage of samples in each stratum. Below this section, FLUX gives the CV that would have been obtained if the samples were optimally distributed among the strata. In the Figure 28, the CV could have been reduced from .130 to .093 with an optimal sample allocation.